DOCUMENT RESUME

ED 186 438 TM 010 035

AUTHOR Hansen, Joe B., Comp.
TITLE Report of the Committee to Examine Issues Related to the Use of the Norm Referenced Model for Title I Evaluation.
INSTITUTION Northwest Regional Educational Lab., Portland, Oreg.
PUB DATE Oct 78
NOTE 99p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Achievement Gains; Advisory Committees; Compensatory Education; Elementary Secondary Education; *Evaluation Methods; *Mathematical Models; *Norm Referenced Tests; *Pretests Posttests; Program Evaluation; Statistical Bias; *Testing Problems; Test Interpretation; *Test Norms; Time
IDENTIFIERS Elementary Secondary Education Act Title I; Equipercentile Assumption; Out of Level Testing; Regression Toward the Mean; *RMC Models; Technical Assistance Centers

ABSTRACT

This report documents the activities of the technical committee formed to examine issues related to the norm-referenced model (RMC Model A) for evaluation of Elementary and Secondary Education Act Title I programs. Background information is provided on the committee's activities, purpose, and meetings; and the members are listed. The reports of individual subcommittees dealing with specific technical issues are presented: (1) equipercentile assumption; (2) use of norms; (3) interpolation and extrapolation; (4) separation of selection from pretest/regression effect; (5) out-of-level testing; and (6) RMC Model A2, the non-normed testing model. These committee reports comprise almost one-third of the document, and consist mainly of conclusions and recommendations, including lists of unresolved issues. Attachment 1 contains Brent Bridgeman's paper, Extrapolation and Interpolation in Model A1 Title I Evaluation; Robin Baker and Keith McNeil's Technical Committee Recommendations Regarding Testing Times in All Models; and Baker and Tom Williams' Issues Related to Interpolation. Attachment 2 contains Stephen L. Murray's Identifying and Controlling Sources of Regression Bias in Model A; a memo by Gene Glass on the regression effect; Brent Bridgeman's Practical Alternative to Model A; and Henry Scherich's A Case for Instructionally Dated Norms. (Author/GDC)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED186438

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

# REPORT OF THE COMMITTEE TO EXAMINE ISSUES RELATED TO THE USE OF THE NORM REFERENCED MODEL FOR TITLE I EVALUATION

October 1978

Compiled by Joe B. Hansen, Director
Title I Evaluation
Technical Assistance Centers
Northwest Regional Educational Laboratory
710 S. W. Second Avenue
Portland, Oregon 97204

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Joe HANSEN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM010 035

TABLE OF CONTENTS


| | | Page |
|---|---|---|
| I. | PURPOSE | 1 |
| | BACKGROUND | 3 |
| II. | COMMITTEE REPORTS | 5 |
| | • Committee on Interpolation and Extrapolation | 5 |
| | • Committee on Use of Norms | 9 |
| | • Committee on Separation of Selection and Pretest | 17 |
| | • Committee on Out-of-Level Testing, Model A2 and the Equipercentile Assumption | 18 |
| | Questions Developed at Full Committee Meeting - Denver | 27 |
| III. | SUPPORTING DOCUMENTS | |
| | Attachment 1 | |
| | • Extrapolation and Interpolation in Model A1 Title I Evaluation. Brent Bridgeman, Educational Testing Service | |
| | • Technical Committee Recommendations Regarding Testing Times in all Models. Robin Baker and Keith McNeil, National Testing Service | |
| | • Issues Related to Interpolation. Robin Baker and Tom Williams, National Testing Service | |
| | Attachment 2 | |
| | • Identifying and Controlling Sources of Regression Bias in Model A. Stephen Murray, NWREL | |
| | • Regression Effect. Gene Glass | |
| | • A Practical Alternative to Model A. Brent Bridgeman, Educational Testing Service | |
| | • A Case for Instructionally Dated Norms. Henry Scherich | |
| | Attachment 3 | |
| | • Equipercentile Committee Recommendations. Steve Murray, NWREL | |

## PURPOSE

The purpose of this report is to describe and document the activities of the technical committees formed to examine issues related to the implementation and use of the Norm-Referenced Model (Model A) for Title I evaluation. The report:

- provides background information on the formation of the committees, their purpose and responsibilities, their mode of operation, the schedule of committee meetings and committee membership,
- contains the reports from each of the individual subcommittees formed to deal with specific technical issues,
- provides supporting documents in the form of technical papers, memos and conference papers (Attachments 1, 2 and 3) related to the various technical issues addressed by the committee.

## BACKGROUND

The Model A Technical Committee is actually a group of subcommittees, each of which was formed to deal with specific issues related to the use of the Norm-Referenced Model. The committee was formed at the request of Dr. Ronald Fishbein, USOE/OPBE, Project Officer for the Title I Evaluation Technical Assistance Centers. The Technical Committee concept resulted as an outgrowth of discussions of technical issues in the quarterly TAC Directors' meetings, wherein various TAC Directors and USOE staff were seeking answers to many thorny problems regarding the Title I Evaluation Models.

In working with SEA and LEA clientele, TAC staff are frequently confronted with difficult issues of both practical and technical natures for which clear and consistent answers have not been established. The TAC Directors had turned to USOE for guidance on these issues. USOE recognized that effective guidance could not be provided without the benefit of a thorough examination of issues by qualified skilled professionals, who themselves were familiar with the issues. They further recognized the existence of a vast resource pool of expertise in the personnel of the various TACs. As the development of regulations for the implementation of Section 151 of Public Law 93-380 creating the evaluation models became imminent, USOE felt the need for expert advice on a variety of issues related to the Evaluation Models. USOE felt that the formation of technical committees from among the TAC personnel with external experts from the fields of measurement and evaluation functioning as consultants would provide the best means of obtaining such advice. Each committee was charged with the responsibility of examining a specific issue or group of related issues and providing recommendations for USOE to use in developing the implementation regulations. Committees were formed for Models A and C and a separate committee was formed for revision of the RMC developed reporting forms and publications. The proceedings of the Model C committee have been summarized in a separate document by Dr. Gary Eckternacht, Director of the Region 2 Technical Assistance Center. This document deals with only the proceedings and recommendations of the committees on Model A, the Norm-Referenced Model. As a result of discussions among the TAC Directors and USOE representatives at the quarterly directors' meeting in January 1978, it was decided that the issues examined by the Model A committee would be:

- Equipercentile Assumption
- The Use of Norms
- Interpolation and Extrapolation
- Separation of Selection from Pretest (Regression Effect)
- Out-of-Level Testing
- Model A2

Each TAC Director was asked to identify people on his or her staff who would have the appropriate expertise to deal with the respective issues. Dr. Fishbein asked Joe Hansen to assemble the committees, chair and coordinate their activities. The following committee assignments resulted.

Committee on Use of Norms

| | |
|---|---|
| Nelson Noggle | TAC Regions 8, 9, 10 |
| Dan Ozenne | USOE/OPBE |
| George Temp | TAC Region 4 |
| Hank Scherich | TAC Region 3 |
| Anthony Allen | TAC Region 1 |

Committee on Selection and Pretest Separation

| | |
|---|---|
| Robin Baker | TAC Region 3 |
| Tom Williams | TAC Region 3 |
| George Temp | TAC Region 4 |
| Steve Murray | TAC Region 10 |
| Steve Horwitz | TAC Region 1 |

Committee on Interpolation and Extrapolation

| | |
|---|---|
| Robin Baker | TAC Region 3 |
| Brent Bridgeman | TAC Region 3 |
| Nelson Noggle | TAC Regions 8, 9, 10 |
| Hank Scherich | TAC Region 3 |
| Tom Williams | TAC Region 3 |

Committee on the Equipercentile Assumption

| | |
|---|---|
| Robin Baker | TAC Region 3 |
| Gary Eckternacht | TAC Region 2 |
| Nelson Noggle | TAC Regions 8, 9, 10 |
| Dan Ozenne | USOE/OPBE |
| Steve Horwitz | TAC Region 1 |

Committee on Out-of-Level Testing

| | |
|---|---|
| Robin Baker | TAC Region 3 |
| Steve Horwitz | TAC Region 1 |
| Diane Proctor | TAC Region 3 |
| Dan Ozenne | USOE/OPBE |
| Nina Selz | TAC Region 6 |

Committee on Model A2

| | |
|---|---|
| Paul Raffeld | TAC Region 6 |
| Nelson Noggle | TAC Regions 8, 9, 10 |
| Mike Bunch | TAC Region 3 |
| Ron Fishbein | USOE/OPBE |

## COMMITTEE REPORTS

### Committee on Interpolation and Extrapolation

This committee first met in Denver on March 14th and 15th in conjunction with the committee on norms and a separation of selection and pretest. Robin Baker and Tom Williams of NTS (Region 3) prepared a brief paper in advance. This paper related to interpolation is part of Attachment 1. The following subcommittee report was prepared by Robin Baker.

#### Technical Committee Recommendations Regarding Various Interpolation/Extrapolation Issues

Recommendation 1. It was the unanimous opinion of the Technical Advisory Committee that extrapolation rather than interpolation be used when a test with fall/spring norms is used and the testing date is either prior to the fall norms or after the spring norms. This issue has been discussed at length in papers by Baker and Williams and Bridgeman. The complications caused by differences in the school year versus summer growth rates as well as the uncertainty about whether the summer should be counted as one or three months led the Committee to this recommendation.

Recommendation 2. It was the unanimous opinion of the Technical Advisory Committee that whenever interpolation/extrapolation needs to be done, expanded standard scores should be used rather than NCEs. The relative merits of using these two metrics have been discussed in a paper by Baker and Williams which concluded that although expanded standard scores were preferable, the ease with which interpolation could be done using NCEs made it the method of choice. Bridgeman, however, has developed a method that uses standard scores but replaces the tedious percentile line bracketing procedure with a much simpler method that rests upon two assumptions: first, that growth as measured by expanded standard scores is uniform from pretest to posttest; and second, that neighboring equipercentile lines are parallel to one another.

The Committee felt that expanded standard scores were preferable as they are the metric that most test publishers use to link the various levels of their tests together. Standard scores should more accurately reflect, therefore, changes that are attributable to growth. NCEs, on the other hand, are tied to the distributions that are created at particular points in time; as such, they are insensitive to changes in these distributions that occur across time. With Bridgeman's procedure now available to simplify the computational chores, the Committee felt that the expanded standard scores interpolation/extrapolation procedure could now be recommended without reservation.

Recommendation 3. The Committee felt that some clarification was necessary regarding the relationship between the need to interpolate/extrapolate and testing times. In order to obtain an accurate no-treatment expectation in Model A, it is preferable that the temporal distance between the testing dates

be the same as temporal distance between the norming dates of the instrument being used. Such a recommendation has been typically stated as, "when pre- and posttesting dates deviate from the norming dates, the deviation should be in the same direction and of approximately the same size" (Users Guide, p. 41).

The Committee felt that this recommendation would best be operationalized in the following manner:

- determine the relative position of the pretest date to the date of the norms being used for the pretest.

- determine the date lying in the same relative position to the norms being used for the posttest. This date along with a plus or minus two-week leeway describes a four week interval during which posttesting should be done. If posttesting is done during this interval, then the pretest/posttest interval can be considered comparable to the interval lying between the pretest and posttest norm dates.

- both the pretest and the posttest must lie within six weeks of the norm dates. If they lie within two weeks of the norm dates, interpolation/extrapolation is unnecessary, and the publishers' norms can be used.

This recommendation is graphically illustrated in Figure 1. A fall/spring testing program is being used with an achievement test that has mid-October and mid-April norms. The pretest has been administered at the beginning of October, exactly two weeks prior to the norm date. This means that in order to maintain the appropriate inter-test interval, one should administer the posttest in the period from four weeks prior to the posttest norm date to exactly at the posttest norm date (e.g., the period from mid-March to mid-April). If the posttesting is done in the period from the beginning of April to mid-April, then there will be no means to interpolate/extrapolate as this falls within two weeks of the norm date. On the other hand, if posttesting is done between mid-March and the beginning of April, then interpolation/extrapolation will be necessary.

Should the posttesting not be done within the recommended four week period, then it was the Committee's opinion that interpolation/extrapolation would be necessary regardless. In Figure 1, if posttesting was done between mid-April and the end of April, interpolation/extrapolation would be necessary even though this date fell within two weeks of the publishers' norms. It was the Committee's opinion that to have a test administered during such a period and scored using the publishers' norms would create a situation where the inter-test interval could be as much as four weeks longer than the inter-norm interval.

As mentioned earlier, this particular recommendation is a clarification of existing information rather than the resolution of some issue. It was the Committee's opinion that both qualifications, having an inter-test interval that is within two weeks of the inter-norm interval and having all testing

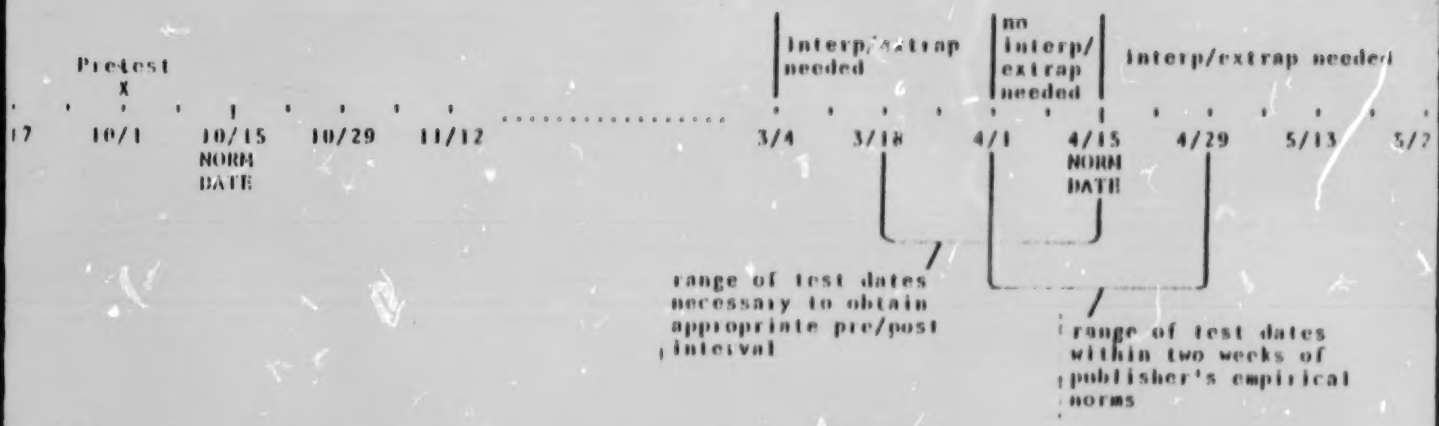

Pretest
X
Interp/extrap needed
no interp/extrap needed
Interp/extrap needed
17
10/1
10/15
NORM DATE
10/29
11/12
3/4
3/18
4/1
4/15
NORM DATE
4/29
5/13
5/27
range of test dates necessary to obtain appropriate pre/post interval
range of test dates within two weeks of publisher's empirical norms

Figure 1

done within two weeks of the norm dates, must be met in order that interpolation/extrapolation is unnecessary. Testing within two weeks of the norm dates does not in and of itself guarantee that interpolation/extrapolation will be unnecessary.

Issues Requiring Study. Several issues arose during the discussions that the Committee felt would require an in-depth examination at some future time. The first issue concerned testing dates and their effect on the aggregatability of the data. Most testing programs typically use one of four schedules: fall/spring, fall/fall, spring/spring, or mid-year/mid-year. At the most elementary level, the issue concerns how should results from a six or seven month testing interval be aggregated with results from a twelve month testing interval. Bridgeman has suggested one obvious solution, namely, to calculate an NCE rate (NCE gain divided by number of months between pre- and posttesting) as the measure to be aggregated. The Committee felt, however, that the issue of aggregation given different testing programs goes much deeper than this. For example, is a mid-year/mid-year testing program going to produce results comparable to a spring/spring testing program? This seems highly unlikely as the mid-year/mid-year testing program is probably reflecting the results of two different school year treatments, while the spring/spring testing program will, in all likelihood, reflect the effects of only one school year program. Both programs, of course, include summer growth which the fall/spring program ignores. Recent research by Pelavin and Stenner et al. indicates that there is a strong possibility that Title I students experience a summer loss. This is reflected in a series of fall/spring evaluations, all of which show strong increments in growth; the sum of these growths, however, far exceeds the observed growth that occurs over the entire period. The gains attributable to the Title I programs are not being maintained after the school year ends. Consequently the questions arise that should Title I programs be evaluated strictly on the gains that they produce (e.g., a fall/spring testing program), or should they be evaluated not only on the growth that they produce but on the maintenance of that growth (e.g., a fall/fall testing program).

Tying to this issue is also the question, should Title I evaluation testing be done to include as much of the program as possible. If we accept this, then a nine month Title I program should be evaluated with a September/May testing program. If the test being used, however, has October/April norms, then the evaluation will be in violation of recommendation number 3. The user will be testing six weeks prior to the fall norming date and six weeks after the spring norming date.

Thus, it was the Committee's recommendation that another committee be brought together to examine this question of testing dates. The Committee suggested that in addition to NCE rates, other measures that might be considered would be the gain observed between two standardized times, the gain observed across the duration of each Title I program, or the gain from one norm date to the next norm date. These were only suggestions, however. It is by no means clear that aggregation of data from different testing programs is possible. If this subsequent committee cannot arrive at a procedure that will permit complete aggregation, then it will need to specify the extent to which results from different programs must be kept separate.

One other issue that the Committee felt needed further exploration was the matter of whether ten or twelve month years should be used in interpolating results from once-a-year testing programs. The Committee suggested that it might be useful to gather information from the publishers regarding the number of months per year that they use on smoothing their growth curves.

With the submission of this report, the subcommittee on Interpolation and Extrapolation Issues concluded its work. Additional information is provided in Attachment 1.

## Committee on Use of Norms

This committee met twice. The first committee meeting took place in Denver on March 14th and 15th. At this initial meeting, several broad issues were identified and a number of questions were raised regarding the use of norms with Model A. Due to a lack of time, the committee chose to limit its task to that of identifying central issues and raising questions. The following report is therefore limited and should be viewed only as a preliminary position statement. The report was prepared by Nelson Noggle.

### Report from the Technical Committee on the Use of Norms in Model A

Central Issue. It is very difficult to express the central issue if one thinks only in terms of measurement practice under Model A; however, it becomes clear if we view it in terms of an appropriate evaluation design: Are national or local norms an adequate source of comparison? Six questions related to this central issue were raised:

1. What can TACs do to foster use of quality norms?
2. Are national norms appropriate?
3. Are local norms better?
4. Are special population norms better?
5. Are the user-based estimates of empirical norms as good as empirically developed norms?
6. What are the norm-related implications of chance-level scores?

Committee Position. The Technical Committee realizes a serious design weakness when national norms are used. While local norms appear to be better, they suffer from practical problems associated with their selection. Special population norms might be an improvement, but there are serious questions of definition and source. The question of SRA and HM user-based norms requires more investigation, but raises implications about whether or not TACs analyze or disseminate the results. Both floor and ceiling effects are seen as causing problems when using norms for certain tests and/or levels of tests.

Recommendations and Questions. What can TACs do to foster use of quality norms?

1. Develop a list of factors to look for when evaluating norms.
2. Collect and list publisher's information regarding each factor.
3. Consider development of minimum qualifications for each factor.
4. Major goal--help LEAs and SEAs to become wiser consumers.

Are national norms appropriate as a source of comparison?

1. Reword RMC guidelines which confuses "representative of" and "represented in".
2. Use national norms as last resort rather than first choice.
3. Continue to evaluate appropriateness in light of better local norms or special population norms decisions.
4. Evaluate variations in norms across publishers.
5. Multiple application designs, Model A vs. Model C, should be studied.

Are local norms better?

1. Yes--they meet more closely with desirable design characteristics.
2. The "contamination" effects of recent Title I treatment should be studied.
3. Multiple application designs, national and local, should be studied.
4. If Model A with local norms is possible, why not Model B?
5. What about small districts (small student N)?

Are special population norms better?

1. Need to study definition of Title I related norms.
2. Need to study "contamination" effects of recent vs. past norms.
3. Would a single set of norms, OE developed, and offered as a point-in-time unchanging reference, be preferable? Possible?

What about the user-based estimates of empirical norms?

1. Probably better than interpolation, but more study needed to establish level of acceptability.

2. Who decides on the level of acceptability?

3. Who disseminates the word on the level of acceptability?

What about chance-level scores?

1. There appears to be differences across publishers that have implications for:

   a. in-level testing
   b. out-of-level testing
   c. the other models as well
   d. further study

2. As publishers produce more levels with narrower content ranges, the future could have more problems--need data on new publications.

3. Need to develop a rationale for situations where many students are at chance-level but obtain fairly significant percentile scores (e.g., 20-40).

4. The ceiling effect, especially for posttests, could be another problem because it depends on range of percentiles.

Basis for Recommendations. The collective experiences and training of the group, as well as an external critique by Gene Glass, formed the major basis for recommendations. The Committee readily senses the need for a dynamic system of evaluation models, and acknowledges the responsibility that TACs must assume in constantly improving LEA and SEA evaluations. The Committee also accepts the need for improving the models in an effort to make the data comparable across projects, or models. The Committee realizes that its effort to study and improve uses of norms in Model A has just begun, and that much more must be done. Given more time, however, it appears clear that Model A needs to be analyzed more clearly in terms of design theory.

Recognizing that it had not completed its task, the Norms Committee met again in Chicago on May 15th and 16th. The following report was prepared by Joe Hansen.

Report on Meeting of TAC Technical Committee on Use of Norms with Model A

The Model A Technical Committee for norming issues met in Chicago on May 15 and 16. The following individuals were in attendance:

| | |
|---|---|
| Anthony Allen | Brent Bridgeman |
| Robin Baker | Nelson Noggle |
| Henry Scherich | Dan Ozenne |
| Tom Williams | Joe Hansen |

Terry Strand from ETS Evanston sat in on the meeting. The following is a summary of the discussion that occurred, the conclusions, and the recommendations of the group.

The committee decided at the beginning that it should set an attainable goal for itself to provide focus and direction for the discussion of norms issues. This led to a discussion of the range of possible types of recommendations that are within the scope of the committee. The committee agreed that recommendations should be practicable. That is, they should help users without requiring great changes to occur in either USOE policy or publishers' practices. The committee then addressed the question of what we expect norms to do for us. Two different points of view emerged from this discussion and were subsequently examined by the committee.

First, it was suggested that norms should provide the basis for obtaining a no-treatment expectation. In this sense, they could be viewed as a proxy control group. This position was identified as the experimental position in the ensuing discussion. Second, norms were viewed as a means of establishing a comparison scale on which the gains of Title I projects could be compared against a national standard. This position came to be identified as the comparative position. The committee adopted the posture that the experimental approach was appropriate for local district evaluation and for answering research questions, whereas the comparative approach was more appropriate for purposes of data aggregation and reporting to Congress. It was determined that the two approaches embodied substantially different methodological concerns. A question was raised by the group as to whether the Title I regulations implied which of these approaches, if either, should be taken in Title I evaluation. These two approaches were then further defined as follows:

Comparative approach. Assessing growth against a standard (norms). When the questions of interest are "How is our project doing relative to other projects similar to it?" they are normative questions and are therefore subsumed by the comparative approach.

Experimental approach. Comparing observed growth against the growth that would have occurred without the (Title I) treatment. If the question of interest is "How would Title I students have done in the absence of Title I program?" then it is a design question, and therefore is subsumed under the experimental approach.

Based on the discussion to this point, the committee arrived at two tentative conclusions:

1. If the purpose of Model A is to establish a no-treatment expectation and assess project effectiveness in terms of that no-treatment expectation, then many questions pertaining to norms as performance standards become moot.

2. If we assume that an experimental model is appropriate, then norms must be viewed as a proxy control group. This led to the conclusion that in order to have an adequate control group, it would be necessary to have norms that are free of the contaminating influence of Title I students.

14

Ideal Norms. The committee, recognizing the need to establish criteria for acceptable norms, established the following three characteristics as being minimum requirements for a set of norms to be used in an experimental approach. The norms should be based on students who:

1. are similar as possible to the Title I group in question,

2. have not had the Title I treatment,

3. have received all other instructional treatments received by the Title I group except the Title I treatment.

Norms were examined from the perspective of these requirements: within school, within district, multidistrict, state, regional and national.

The following observations were made in conducting this analysis:

1. Norms beyond the district level should be based on students who match the Title I students on as many demographic variables as possible.

2. The inclusion of non-Title I students at the same level of achievement as the Title I students would likely introduce uncontrollable bias into the norms under the experimental approach.

3. The sustaining effects study has established a cumulative data base which could be used to develop norms for Title I students under a comparative approach.

4. The sustaining effects study and the ESAA study both encountered serious problems in their efforts to establish valid control groups, finding that all students like those in the treatment group were receiving some form of treatment similar to that given the treatment group students. This tends to validate the committee's concern regarding the use of norms as proxy control groups under the experimental approach.

5. Historically-based norms would become suspect due to population or Title I program shifts, or due to changes in Title I definitions.

Based on its analysis, the committee arrived at the conclusion that there is no satisfactory way of constructing a norm (control) group for obtaining a valid no-treatment expectation for Title I students. The committee also recognized that Models B and C have problems of varying kinds of comparison group bias. The committee then shifted its focus to examining the feasibility of a comparative approach to Title I evaluation, and began to examine the possibility of establishing national Title I norms against which Title I projects gains could be compared. It was recognized that such norms would not provide a no-treatment expectation, but might provide the basis for a "Title I growth expectation," which would be defined as the average growth from pretest to posttest for a group of Title I students at a given pretest percentile

level. This would include growth due to maturation, schooling and Title I. An extensive discussion of the possible utility of Title I norms followed. The more salient points in that discussion were:

1. While such a set of norms would be useful in determining the growth of a Title I project relative to the average growth for Title I students, it would not be useful in determining how much growth students in the project would have experienced in the absence of Title I (i.e., the no-treatment expectation).

2. An average Title I effect would be built into the norms. Therefore, if Title I had the effect of raising achievement nationally, new norms would be required each year.

Purposes of Data. The committee then addressed the question of what kinds of data were being asked for in Title I evaluation and for what purposes. Three points emerged from this discussion:

1. Congress seems to want comparative data (football scores).

2. Local districts seem to want comparative data, to answer the question, "How is my Title I project doing relative to other similar Title I projects?"

3. Evaluators, researchers and only a few districts seem to want to answer the question of how Title I students are doing relative to what they would be expected to do in the absence of the Title I program.

The committee then raised several questions regarding the tenability of the experimental approach as a means of providing Congress with the data they need. As the committee grappled with the above issue of multiple uses of data, a realization emerged that the different evaluation purposes needed to be served by different types of data.

Two conclusions emerged:

1. For the overall purpose of aggregating data at the state or national level, a comparative approach would be more consistent.

2. An experimental approach would be desirable at the local level as an option for answering questions about the efficacy of or comparisons among programs.

It was observed that the validity of the comparative approach does not necessarily depend on the validity of the equipercentile assumption, because Title I growth norms established for these purposes could be updated with fresh data each year, making the actual Title I growth patterns observable.

In the discussion of Title I norms and how they should be established, the following questions were addressed:

1. Should national Title I norms be based on Title I students or on Title I projects?

2. Should the norms be based on available district data or on a rationally-drawn sample?

3. Should there be separate norms for each publishers' edition?

4. Is independent norming necessary, or could publishers equate to a common norm; e.g., CTBS as used in the SES study?

5. Should Title I norms reflect both pretest and posttest values or should they be based on a distribution of gains?

6. If Title I growth norms were in terms of NCE units, would computation and interpretation be facilitated?

7. How frequently should growth norms be updated?

The committee's position on each of these issues follows:

1. Title I growth norms should be based on students' rather than project gains. The rationale for this is that many project gains would be based on very small N's and would therefore be expected to vary widely and to be subject to considerable error. In situations where a project is school-wide or district-wide, conclusions based on project norms could not be generalized validly across either projects or schools, but if norms were based on students, generalizations could be made across all Title I students.

2. Title I growth norms could either be based on existing data or developed independently. If the norms are to be based on existing data, three concerns arise. First, is there a sufficient data base in existence? Second, if there is not an adequate data base, how might one be established? Third, when and how should the norms be updated?

   The committee felt that if it became necessary to develop a data base it should be done by a contractor under USOE's sponsorship. An alternative to development would, of course, be a sampling of user data from the current system or from the Sustaining Effects Study.

   Some advantages were seen to using the existing reporting system to establish Title I norms. First, it would not be necessary to administer a single "anchor test" nationwide. Such a test, if it were a publisher's test, might imply an endorsement by USOE. Conversely, if it were a USOE developed test it might imply to some users the advocacy of a national curriculum.

3. Separate norms for each publisher's edition would be too costly. Problems could also arise from the different proportions of Title I students in publisher's norming samples. Separate norms are, therefore, not recommended.

4. Equating would be necessary only if:

   a. pre-post norms were being developed, and
   b. the norms were to be interpreted into the publisher's system.

   If growth norms were established for comparative purposes only, as we advocate, no equating would be necessary.

5. Since achievement gain under Title I is the greatest concern, the Title I norms should be growth norms rather than fall, mid-year or spring norms. Pretest-posttest data are not of interest in themselves at the aggregation level.

6. Having Title I growth norms expressed in NCE units rather than standard scores or percentiles would facilitate computation by reducing the number of table lookups and conversions required of the user. Computation would be further facilitated by having separate norms tables for different instructional intervals, such as October to May or May to May. If the growth norms were in NCE units, aggregation of project gains would be more easily accomplished.

7. The growth norms should be updated annually. Updating could occur as a function of the annual reporting cycle. This would make it possible to compare each successive year's growth with the previous year and could provide a basis for longitudinal evaluations. The committee also viewed the proposed growth norms as having additional utility for LEA use of the evaluation data. The LEA would be able to compare to the following growth statistics (norms):

   a. LEA past
   b. SEA present
   c. SEA past
   d. Nationwide present
   e. Nationwide past
   f. Subcategories present
   g. Subcategories past

Conclusion. The committee recognizes that Model A as currently defined does not yield a valid no-treatment expectation. It is also recognized that approximately 95 percent of all local districts using one of the models will prefer to use Model A because of the relative ease of implementation when compared with Models B and C. What we propose is a dual system wherein a comparative approach is employed for purposes of aggregation and an experimental approach is used for a more definitive measure of local project effectiveness. This would not require any changes in the implementation procedures currently proposed for Model A. The TAC's role in supporting this

modified system would remain essentially the same as in the current system, providing training and consultation to LEAs to assist them in meeting the federal reporting requirements and to improve local programs through more effective evaluation methodology.

## Committee on Separation of Selection and Pretest

This committee met in Denver on March 14th and 15th. Prior to the meeting, Steve Murray dedicated considerable effort and time to the development of a paper demonstrating that the separation of pretest and selection test procedures did not adequately control for the regression effect. A copy of this paper is included as part of Attachment 2. Dr. Gene Glass of the University of Colorado, Education Research Laboratory, reviewed Murray's paper and commented upon it in a memo addressed to all interested parties. Dr. Glass, who had been contracted as a consultant, met with the group in Denver. His memo (a portion of Attachment 2) follows the paper by Murray. The following report is the result of the Denver meeting of this committee. As was the case with the initial meeting of the Norms Committee, this committee chose to identify a central issue and establish a committee position with respect to the issue. Unlike the Norms Committee in its first meeting, this committee felt there was sufficient evidence to warrant specific recommendations regarding the issues identified. The following report was prepared by Joe Hansen, Steve Murray and Nelson Noggle.

### Recommendations from the Technical Committee on Selection and Pretesting

Central Issue. Does separating the selection test from the pretest have the properties claimed for it in the information disseminated through the operation of the Title I Technical Assistance Centers?

Committee Position. The Technical Committee seriously questions the soundness of the claim that separating the selection test from the pretest eliminates regression to the mean in the Norm Referenced Model. The committee believes that the recommendations which follow will improve the technical merit of the Norm Referenced Evaluation Model and avoid serious misunderstanding over the requirements for sound implementation.

Recommendations.

1. Hold publication and dissemination of RMC's Technical Bulletin #3, The Regression Effect.

2. Separating the selection test from the pretest should no longer be implied as an adequate control for regression to the mean.

   a. Title I evaluation implementation regulations (rules?) should exclude the requirement that the selection test and the pretest must be separated.

b. Regardless of whether the pretest or a separate test is used for selection, an appropriate statistical adjustment is required for estimating the no-treatment expectation. The adjustment formula would be different for each case.

3. A simulation study should be conducted to demonstrate the impact of alternative procedures of estimating the no-treatment expectation. This study would include the present selection and pretest design as the sole control for regression to the mean, as well as suggested alternatives plus corrections.

4. Consideration should be given to generating tables of correction factors (e.g., test-retest reliabilities) and procedures for using these tables in generating no-treatment expectations for the Norm Referenced Evaluation Model.

5. Test publishers should be advised that implementation of the Title I Norm Referenced Evaluation Model will require test-retest correlations, cross level correlations, and that alternate form correlations may be required depending upon alternative evaluation designs actually implemented.

Basis for Recommendations. The Technical Committee evaluated and discussed papers submitted by Steve Murray and Gene Glass. Dr. Murray's paper indicates through an extension of Kenny's analysis of statistical adjustment techniques for the nonequivalent control group design that regression exists beyond that which can be controlled solely by a separate pretest. Dr. Glass's paper indicates that Cronbach's "generalizability theory" supplants the classical notion that errors of measurement are intrinsic to a test, and that reliability is a property of a particular use of a test. Dr. Glass and Dr. John R. Hills reviewed Dr. Murray's paper, and have agreed with its findings.

The papers by Murray, Glass and Bridgeman (Attachment 2) are provided as support for the Technical Committee recommendations.

The final meeting of the Model A Technical Committee took place in Denver on June 15 and 16. Three subcommittees were formed to discuss the equipercentile assumption, out-of-level testing and Model A2. The following reports are the results of that meeting. The report on Out-of-Level Testing was prepared by Joe Hansen. The report on Model A2 was prepared by Paul Raffeld and the report on the Equipercentile Assumption was prepared by Bob Burton.

## Report of the Committees on Out-of-Level Testing, Model A2 and the Equipercentile Assumption

The Model A Technical Committees on the equipercentile assumption level testing and Model A2 met in Denver on June 15 and 16. The following persons were in attendance: Mike Bunch, Bob Burton, Gary Eckternacht, Ron Fishbein, Joe Hansen, Phil Harvey, Steve Horwitz, Steve Murray, Nelson Noggle, Dan Ozenne, Diane Proctor, Paul Raffeld, Nina Selz, Bob Stonehill and George Temp.

The meeting began with a statement of purpose by Joe Hansen, committee chairperson and a discussion of a strategy for accomplishing the tasks. The purpose as stated, was to identify and examine issues of both a theoretical and a practical nature underlying Model A and to make recommendations to USOE regarding (1) issues to be examined further by the issues analysis contractor, and (2) changes which might be considered in the implementation rules for Model A.

The discussion began with an attempt to arrive at a suitable working strategy for dealing with problems. Four different strategies were considered:

1. Work as a large intact group to discuss issues in each of the three areas of the equipercentile assumption, level testing, and Model A2.

2. Divide into three separate work groups with members self-selecting into respective groups based on interest.

3. Break into two separate groups to discuss equipercentile and Model A2 then combine to discuss out-of-level testing and to summarize.

4. Begin as a combined group, identify issues to be pursued by individual work groups then break into three work groups and reconvene for summary.

After much discussion, the fourth alternative was finally agreed upon as the most acceptable. The full committee then addressed each of the three problem areas identifying specific questions of interest within each area, for consideration by the work groups or subcommittees.

## Subcommittee Reports

On Friday morning, the three working groups reconvened after having discussed issues within their respective areas of interest and an individual from each group reported the recommendations to the entire committee. Numerous questions were asked and there was much good general discussion of issues. Following this discussion, an individual from each of the three groups was identified to accept the responsibility for providing a written subcommittee report to the chairman. The subcommittee reports for each of the groups follow.

## Subcommittee on Out-of-Level Testing

This subcommittee was comprised of the following individuals: Joe Hansen, Chairperson, George Temp, Phil Harvey, Steve Horwitz, Nina Selz, Diane Proctor, Dan Ozenne. The committee began by trying to develop answers for each of the questions identified earlier in the full committee meeting (see list on pages 27-28). After several hours, during which the committee had generated answers to only three questions, it became obvious that this was a fruitless strategy. It was therefore decided that rather than trying to answer each of these questions in detail, the committee would instead attempt

to surface issues and make recommendations including hypothesis, methods and resources for USOE to refer to the issues analysis contractor. The following is a summary of the recommendations and some of the relevant points of discussion pertaining to those recommendations.

1. The issues analysis contractor should: (a) examine data from the sustaining effects study including out-of-level norms for the CTBS (other data bases such as ETS, CTB/McGraw-Hill and GATES should also be examined); (b) engage in theoretical work to determine the amount of error in vertical equating and how error is related to test width (the paper by Chris Wood of RMC on norming was identified as a useful resource for this activity); and (c) examine problems/implications for out-of-level testing.

2. The contractor should conduct validity studies to determine whether in-level or out-of-level scores are valid. The discussion regarding this recommendation focused on the value and the interpretation of out-of-level standard scores.

3. There are many alternative ways of determining floor and ceiling effects. The new contractor should examine the body of knowledge pertaining to floor and ceiling effects and ascertain whether RMC's rule of thumb that the group mean should be greater than one-third but less than three-fourths of the total number of items is indeed the best way of insuring against a floor or ceiling effect.

4. The contractor should examine the assumptions and the technical underpinnings of out-of-level testing. In other words, is there a good technical basis for out-of-level testing? It was noted that there are many different ways of deciding when to test out-of-level, all of which are related to different methods of determining floor and ceiling effects. RMC recommends only one way. Lord (1952) gives a method based on optimum P-values (Psychometrica, 1952, 17, 181-193).

5. The issues analysis contractor should examine the body of lore and knowledge regarding different methods of determining floor and ceiling effects and make recommendations to both publishers and TACs regarding these various methods.

6. The committee agreed that if a particular TAC is using methods of determining floor and ceiling effects that are different than those recommended by RMC, it should share those methods with all TACs.

7. The committee agreed that if a test level is too difficult, but going to a lower level would result in inappropriate content coverage, then an alternative test and not a lower level should be considered.

8. After some discussion of the need for guidelines for level selection, the committee offers the following suggestions:

   a. Use of Dolch or Spache lists
   b. P-values equal to .75

With respect to the advantages and disadvantages of out-of-level testing, the committee offers the following observations:

1. Out-of-level testing should provide a more accurate measure of ability and should not be represented to LEAs as a method by which gains can be increased.

2. Out-of-level testing should reduce the student frustration resulting from a test with a reading level that is too difficult.

3. There is a disadvantage to out-of-level testing posed by the problems encountered in trying to convert out-of-level standard scores to in-level percentile scores if the test that was administered is too many levels below the recommended level.

4. An out-of-level mean when translated to an in-level percentile may result in a less accurate measure than if the median is used where a floor effect exists with an in-level test.

The committee addressed the question of what problems are associated with wider versus narrower level spans. Two problems were identified:

1. It might be necessary to have different rules concerning how far out-of-level you might go with narrow levels versus wide levels.

2. Aggregating out-of-level data from both wide band and narrow band tests may be inappropriate. A single level per grade may result in more accurate NCE scores. A question here is whether we want to recommend that when faced with a choice, the LEA select a test with a single level per grade. Such a recommendation, it is recognized, could have legal implications since it will limit the LEAs' choices with respect to publishers.

In trying to identify which of the many issues pertaining to out-of-level testing were most important, the following questions and points of discussion arose. George Temp suggested that there is a single overriding issue. "Is out-of-level testing technically sound?" If we can answer that question, we can handle all others. The committee then tried to identify more precisely what was meant by "technically sound." Two major questions were identified as being pertinent to this issue: (1) Are the equating procedures used by publishers acceptable? (2) Is it conceptually defensible to convert out-of-level percentiles? It was felt by most that the answer to this was no, unless both levels had been given to the same population sample.

A question arose within the committee as to whether it was appropriate to change test levels between the pretest and the posttest. In other words, if the pretest results in a ceiling effect, should a lower level as the posttest be administered? It was agreed that the most important consideration is that the level of test administered be appropriate at the point in time at which it is administered. Related to this, was the question of which type of error is

worse, content or ceiling effect? The committee arrived at two recommendations regarding this issue:

1. Testing should be done at functional level both pre and post. If an inappropriate level is used on the pretest, then an appropriate level should be used on the posttest.

2. Care must be taken in level selection to make sure that the initial test level is correct, then administer the same level on the posttest. It is fair to say that the committee did not have consensus on this recommendation. Steve Horwitz pointed out that it is not always possible to tell what level of a test is appropriate unless you administer a test first. He seemed to prefer to have a pretest ceiling effect and be able to show some gain rather than to have major content or construct departures between the test level and the curriculum.

3. The committee further recommends that the contractor compile a series of case studies based on actual data from the field, to determine the effects of out-of-level testing.

4. A study should be conducted by the issues analysis contractor comparing results between on-level and off-level testing for the same group of students.

The committee concluded that out-of-level testing seems to be technically defensible under certain conditions. However, we need more data to understand how it works and what criteria should be used for selecting an appropriate level. The committee seemed to agree that the use of out-of-level testing would result in increased precision of estimates of NCE gain if and only if out-of-level testing were done correctly.

Conclusion/Summary. The committee agreed on the three general statements of principle that follow:

1. The RMC Guidelines for out-of-level testing are satisfactory for the time being. George Temp, however, feels that the guidelines should be broadened somewhat to allow departures by as much as two or three or even more levels until we have better data to indicate what problems this poses.

2. Questions of how far out-of-level to go must be test specific. More data are needed on various publishers' tests. Generalizations across tests may not be possible.

3. The further out-of-level you test the more caution must be used in interpreting results.

Technical Recommendations Concerning the "Equipercentile Assumption" of Model A1

The following recommendations are based on the meetings in Denver on June 15-16, 1978. They represent an attempt to blend subcommittee recommendations with compromise positions reached during full committee discussion.

The purpose of the recommendations is not to attack Model A1 as an invalid practical tool for evaluation. It is rather to provide a firmer ground, both theoretically and empirically, for use of Model A1. Without such firmer ground, it will never be possible to assess the appropriateness of the model in a given situation.

1. The underlying assumption of Model A1 should be explicitly stated in the User's Guide and in other RMC materials. This assumption is: the no-treatment expectation for a Title I group is the mean NCE score obtained by that group at pretest time.

   The assumption can also be stated as follows: in the absence of any special treatment, the mean NCE score of a group is expected to remain constant over time.

2. The phrase "equipercentile assumption" should be avoided whenever possible. The Model A1 assumption is more properly an "equi-NCE assumption."

3. Alternatively, an "equi-expanded-scale-score assumption" can be used, since computations can be performed in this metric as well as in NCEs.

4. There is currently no mathematical model for which the "Model" A1 assumption holds. Theoretical work is necessary to develop models for which (a) scores have a multivariate normal distribution, and (b) there is no regression effect from test administration #2 to test administration #3. In the absence of such a model, it is impossible to determine if and when A1 is appropriate.

5. Theoretical work on a mathematical model should focus on at least three considerations. First, for a given Title I treatment group, what is the mean toward which scores are regressing? Second, and related to the first consideration, can a given Title I treatment group be thought of as a sample from a normally distributed population? And third, is it possible that a regression effect is masked by a countervailing tendency for Title I students to lose ground over time in relative standing?

6. It is recommended that an empirical study be designed and implemented to test whatever mathematical model(s) are developed. This study might involve secondary analysis of data or new data collection or both.

7. Both theoretical and empirical work should consider trivariate score distributions, so that separation of selection and pretest can be further examined.

Steve Murray has provided some observations and comments on the recommendations of the equipercentile committee. His comments can be found in the memo in Attachment 3.

Subcommittee on Model A2

The subcommittee made the following recommendations regarding the implementation of a non-normed testing model.

1. The TACs should provide for a hierarchy of strategies for analyzing non-normed test data that will permit LEAs of varying size and technical sophistication to select a particular strategy.

   A tentative hierarchy was proposed by the subcommittee:

   a. Develop normalized test score distributions for the non-normed and normed tests at the local level. (This requires a relatively large LEA and a fairly high level of technical expertise or a lot of help from the TACs.)
   b. Use a curvilinear analogue for equating tests across the entire score range as described in Angoff (1971).
   c. Use a linear analog for equating tests as described in Angoff (1971).
   d. Use the current A2 strategy. However, if this alternative is chosen, it is recommended that the TACs make efforts to assist LEAs in: (1) the development or selection of non-normed tests, and (2) the selection of an appropriate level of a normed test in order to avoid the possible effects of score range distribution.

2. The subcommittee made a general recommendation that all TACs provide assistance to LEAs in: (a) strategy selection (selecting the strategy that meets their needs as well as one that the LEA can handle and understand); (b) non-normed test development or selection to meet the requirements of summative evaluation; and (c) normed test level selection in order to ensure more valid test scores with a reasonable score range.

   In short the TACs should take responsibility for the preventative and curative aspects of using non-normed tests for national aggregation and should systematically share with each other and the LEAs any new strategy developments.

3. TACs, LEAs, and the issues analysis contractor should begin to investigate the potential of Latent Trait models for equating tests.

4. The subcommittee felt that in most instances the term equating was not appropriate since "true equating" was not possible with non-parallel tests. The term "estimating" was recommended instead. That is, we estimate the result of a normed test from the information obtained from a non-normed test.

5. Regarding an appropriate sample size, the subcommittee recommended that a minimum of 30 be used as a guide in order to minimize the combined errors of the two tests and consequently provide more stable estimates.

   While sample sizes less than 30 may be adequate for national aggregation, LEAs should be made aware of the problem in interpreting such data at the local level.

6. A side question for the issues analysis contractor is: How small can the sample size be and still meaningfully use the data for national aggregation?

7. It was recommended that one TAC take on the responsibility for developing a "preliminary" set of guidelines for non-normed tests. After the preliminary guidelines are drafted, a committee could be formed to finalize the effort and to ensure representation of all TACs.

   The guidelines should focus on the desirable characteristics of a non-normed test and take a "how to do it" approach to test development and/or selection. The guidelines should be sensitive to local needs and constraints and provide LEAs and TACs with the direction necessary to obtain a reasonably valid non-normed test.

   Some of the characteristics that might be included are:

   a. score range restriction
   b. correlation
   c. reliability
   d. raw score versus mastery score
   e. sensitivity to individualized instruction
   f. race, sex, cultural, and other biases.

8. Apparently, there has been some interest in using Model A2 instead of going out-of-level when one is forced to go several levels out.

   The subcommittee noted that Model A2 does not eliminate the problem, since a normed test is still required for aggregation. Furthermore, test selection should be based on content rather than on the inadequacy of a model.

9. A tentative recommendation by the subcommittee is to avoid equating (estimating) at posttest time since the treatment effect is confounded in the score distribution. That is, the treatment effect would be included in the equating process.

One alternative, however, is to use the spring testing as both a pre- and a posttest and equate only for the pretest. Of course, this would require a broad range of item difficulties.

The subcommittee felt that the issue of equating at posttest time should be turned over to the issues analysis contractor. In addition, the I.A.C. might investigate the minimum correlation required between tests for equating purposes as well as the minimum sample size.

10. Two final issues for the I.A.C. are:

    a. to determine the minimally acceptable method (strategy) to produce aggregatable data using a non-normed test;
    b. to determine if equating and conversion to NCEs mask the real effect. That is, does the procedure eliminate the desirable attributes of a non-normed test?

## Questions Developed at Full Committee Meeting

The following questions were developed by the full committee in the June 15-16 meeting and are reported here for the benefit of the Issues Analysis Contractor and USOE.

### Level Testing

1. How can you decide when to test out-of-level?
2. What are advantages/disadvantages of out-of-level testing within a group? How do you define it?
3. What are problems associated with wider versus narrower level spans?
4. What are practical guidelines for level selection - especially with data?
5. Should we average SS or NCEs?
6. What implications does vertical equating have for out-of-level testing?
7. Reconcile-reiterate previous work on these issues? Is there a (coherent) body of information on these issues?
8. Is there a body of data from publishers? Elsewhere?
9. What are implications of out-of-level norms versus vertical equating? Which do we recommend?
10. How many levels can you go without experiencing serious construct shifts? What is first decision point? Where are tradeoffs? How to establish priorities?
11. Identify assumptions in recommending out-of-level testing?
12. How much difference does it make to test out-of-level? One level? Two levels? Pre-post versus gain?
13. Are there different concerns fall-to-spring than year-to-year?
14. Can you get so far out-of-level that you can't get back in-level?

Equipercentile Issues

1. Is equipercentile assumption tenable?

2. What do joint percentile distributions look like? Is the equipercentile assumption valid under non-normal conditions?

3. Is it possible to state equipercentile assumption in terms that are verifiable? (What is relationship between equipercentile assumption and equal NCE assumption?)

4. Are T-1 students at a given percentile the same as their counterparts in national norms? (Are growth rates the same?)

5. What are the implications inherent in stating Model A - in terms of single versus multiple pop.? (What is the mean?)

6. If not valid--what is recommended?

With the conclusion of the Denver meeting, the Model A Technical Committees completed their assigned tasks. The recommendations and reports contained in this summary are hereby provided for and submitted to USOE/OPBE and RMC Research Corporation for their consideration in pursuing solutions to the issues addressed.

Attachment 1

# Extrapolation and Interpolation in Model A1 Title I Evaluation

Brent Bridgeman
Educational Testing Service

January, 1978

Extrapolation and Interpolation in Model A1 Title I Evaluation

## Introduction

RMC has developed a number of models designed to assess cognitive achievement gains in basic skills projects funded by Title I of the Elementary and Secondary Education Act. In each of the models, "the treatment effect is defined as the difference between observed post-treatment performance (posttest scores) and an estimate of what that performance would have been without the Title I project" (Talmade and Wood, 1976, p. 2).

In Model A1, which is the model most school systems are likely to use, the Title I group is pre- and posttested with the same level and form of a nationally normed achievement test. The no-treatment expectation is that the children in the Title I project will maintain the same percentile rank from pretest to posttest. Thus, for example, if a Title I group were at the fifteenth percentile on the pretest (fall national norms) and at the twentieth percentile on the posttest (spring national norms), the five percentile point gain could be attributed to the Title I project. Gains are actually recorded in standard score units[1] in order to avoid the unequal intervals inherent in percentile scores.

RMC recommends that Model A1 testing be accomplished within two weeks of the midpoint of the normative data collection period. With linear interpolation, testing may be as much as six weeks[2] from the midpoint. It is recommended

---

[1] The standard scores are called Normal Curve Equivalents (NCEs) and have a mean of 50 and a standard deviation of 21.06.

[2] Recent communications from RMC indicate that this may be extended to nine weeks for fall to spring evaluations.

that "when pre- and posttesting dates deviate from the norming dates, the deviations should be in the same direction and of approximately the same size" (Talmadge and Wood, 1976, p. 41). RMC does not recommend a procedure to accomplish this interpolation. The purpose of this paper is to explore procedures for interpolation and extrapolation when testing in Model A1 is not accomplished within two weeks of the normative data collection midpoint. There is no single procedure for interpolation or extrapolation that fits all situations, rather these are options that can be selected to best fit particular assumptions. These options can be best explained with specific examples.

Fall to Spring Testing. Consider a third-grade mathematics project evaluated with the Total Math score from the Metropolitan Achievement Test (MAT). Assume both pre- and posttests were administered six weeks after the midpoint of each normative period. These midpoints for the MAT are October 15 and April 15; thus in this hypothetical example pretesting was on December 1 with posttesting on June 1. The mean pretest standard score for the children in the Title I project was 53 and their posttest score was 65. What is the appropriate percentile equivalent for the pretest score of 53? The answer to this question depends on when the project actually started.

Project start prior to pretesting. Suppose that the project was fully operational by October 15 but that because of scheduling problems testing could not occur until December 1. It would then appear to be most reasonable to estimate the mean national percentile rank of the students when the project began in mid-October. To make this estimate we assume that the growth rate (in standard score units) is the same from October 15 to December 1 as the growth rate from December 1 to June 1. Since the Title I group gained 12 standard score points

in six months (i.e., a rate of 2 points per month) it is assumed that the group would gain 3 points in the 1.5 months from the beginning of the project to the pretest. Thus, our best guess is that on October 15 the project participants had a mean score of 50 (53-3=50) which corresponds to the 12th percentile (NCE=25.3). This extrapolation of the project growth rate is represented by the short dashes in Figure 1.

- - - - - - - - - - - - - - - -

Insert Figure 1 about here

- - - - - - - - - - - - - - - -

Project start delayed. But what if the testing was delayed until December 1 because the project itself was not fully operational before that time? Then it would appear to be most desirable to determine which national percentile would correspond to a standard score of 53 on December 1. Even assuming linear growth over the school year, there is no easy way to quickly estimate the appropriate percentile rank. This is a problem because the expected growth rates of children at different percentiles are not identical. For example, the expected growth in third grade for a child at the 10th percentile on the Total Math score of the MAT is from a standard score of 49 on October 15 to a standard score of 55 on April 15 (i.e., 1 standard score unit per month), while a child at the 40th percentile goes from a 58 to a 68 in the same period (i.e., 1.67 standard score units per month). In order to avoid a time-consuming trial-and-error procedure, a satisfactorially accurate estimate could be made by using the growth rate of the percentile rank identified by the above method (in this example, the 12th percentile). In the 6 months between the empirical norming points the 12th percentile goes from a standard score of 50 to a standard score of 57 (ie., a rate of 1.17 units per month, or 1.75 units from October 15 to December 1).

Subtracting 1.75 from the December 1 pretest mean of 53 yields an estimated October 15 standard score of 51.25, which corresponds to the 15th percentile (NCE=28.2). Thus, we estimate that on December 1 the performance of the Title I group corresponds to the 15th percentile nationally (see dash-dot line on Figure 1).

Posttest after end of program. If the program actually ended around April 15, or if testing were unavoidably delayed in one school but it was desired to make comparisons with other schools that tested near the empirical norming date, then it would be most appropriate to determine the percentile rank of the Title I group on April 15. Given the 2 standard score units per month growth rate in this example, we would estimate that the Title I group would have had a mean standard score of 62 (23rd percentile; NCE=34.4) had they been tested on April 15.

Program end in June. If the June 1 testing date were selected to more closely coincide with the end of the project, then one would want to estimate which percentile rank corresponds to a standard score of 65 on June 1. Again, it is necessary to estimate the rate of growth for children in the appropriate percentile. As previously, we use the percentile from the above method (23rd percentile) to estimate the expected growth rate at this achievement level. The 23rd percentile corresponds to a standard score of 53.7 at the beginning of third grade (Oct. 15) and a standard score of 62 near the end of the third grade (April 15), for a growth rate of about 1.4 standard score units per month, or 2.1 units from April 15 to June 1. Subtracting 2.1 from 65 yields a standard score of 62.9 which corresponds to a precentile rank of 25.7 (NCE = 36.3). These posttest estimates are graphically represented on the right side of Figure 1.

Differences in NCE gains. Since the different methods of interpolating or extrapolating assume different starting and/or ending dates for a project, they must result in different estimates of NCE gain. For the current example, the smallest estimate of gain would result from assuming that the project actually began on December 1 and was completed by April 15. In that time the Title I group went from the 15th to the 23rd percentile, for an NCE gain of 6.2. Assuming that the project actually began on October 15 (12th percentile) and concluded on June 1 (25.7th percentile) results in an NCE gain of 11.0.

Projects which actually ran for the whole school year but pre- and posttested on empirical norming dates that were only six months apart would demonstrate less gain than projects that were no more effective but which used extrapolation to pretest before the fall empirical norming period and/or posttest after the spring norming periods (or that used models B or C to pre- and posttest closer to the actual beginning and end of the program). NCE gain per month (NCE gain divided by the number of months between pretest and posttest) might be a better index of project effectiveness in such instances.

Fall to spring extrapolation or spring to fall interpolation. Instead of of extrapolating the October to April norms to June 1 it would be possible to interpolate the April to the following October norms (in this example, from "end of third" to "beginning of fourth"). However, this period includes the summer months when growth rates may be much slower. These fall to spring to fall growth rates in reading and math (based on national norms for the MAT)

are presented in Figures 2 and 3.[3] Although the Total Reading growth rates

---

Insert Figures 2 and 3 about here

---

over the six months that include the summer are often identical or nearly identical to the school year rates, the growth rates for Total Math are generally substantially lower over the summer months. Using spring to fall norms for the posttest scores in the current example, yields a somewhat different estimate of the June 1 percentile rank. From April 15 to the following October 15 the 23rd percentile Total Math score go from a standard score of 62 to a standard score of 64.5, or about .4 standard score units per

---

[3] An important serendipitous discovery noted in the construction of Figure 2 was the zero growth expectation in Total Reading for fourth percentile children in the third grade. It seems highly unlikely that there is no real growth for these children in six months of third grade. The problem appears to be a floor effect for the MAT norm group at the beginning of third grade where children at the fourth percentile have a mean raw score of 26 (compared to a chance score of 25). For the norm group as a whole there is also evidence of a floor effect for the Total Reading score at the beginning of third grade. The mean raw score is about 61 (SD = 18) with a median score of 67. This one-third of a standard deviation difference between the mean and median indicates a substantial positive skew and a probable floor effect. End of third-grade norms are based on the administration of a different test (Elementary instead of Primary II) and the performance of the fourth percentile is less than chance (mean raw score of 20 compared to a chance raw score of 23.75). Since the problem is in the testing of the normative sample itself, out-of-level testing of Title I group in a Model A evaluation would not solve the problem. It would be totally inappropriate for a third-grade Title I reading program that had pretest mean scores below or about the 10th percentile to use the MAT in a Model A evaluation. Since other tests at this or other grade levels might suffer from similar deficencies, it is strongly recommended that whenever pretest performance is below the 10th percentile, the growth expectations based on national norms be closely examined.

month. Projecting this growth rate back from the posttest standard score of 65 yields a standard score estimate of 64.4 on April 15, which corresponds to a percentile rank of 29.6 (NCE = 38.7). Thus, in this example the difference between extrapolating fall to spring norms and interpolating spring to fall norms is about 2.4 NCEs. Inspection of Figures 2 and 3 suggest that, at least for the MAT, differences between interpolation and extrapolation would generally be of about this size or smaller, except for second grade programs where "school year" growth rates are substantially higher than "summer" growth rates. A second grade math program with a mean standard score of 54.75 on a June 1 posttest would be at the 20th percentile (NCE = 32.3) based on an extrapolation of fall to spring norms, and at about the 35th percentile (NCE = 41.9) based on an interpolation of spring to fall norms. Since this difference of 9.6 NCEs could significantly alter judgments of program effectiveness, comparable results would be possible only if a consistent national policy is adopted.

In the absence of empirical data on actual month-by-month growth rates, the choice between extrapolation and interpolation must be made on logical grounds. Recognizing that growth during the school year is not uniform, it would still seem more reasonable to estimate growth during a school year program from school year norms than from norms that included three summer months. For testing after the fall normative period or before the spring normative period, norms should be interpolated, and norms should be extrapolated for testing before the fall empirical norming date or after the spring empirical norming date.

Year to Year Testing

Interpolation and extrapolation is less precise for projects that use a test that has only one empirical norming point during the year. In order to either interpolate or extrapolate some measure of growth rate over the school year is needed. As indicated above in Figures 2 and 3, growth rates over the summer months are usually, but not always, slower than growth rates during the school year. Furthermore, differences between school year and summer growth rates may vary from one academic subject area to another and from one grade level to another. One assumption that is frequently used by test publishers is that growth over the three summer months is equal to one month's growth during the school year, or to put it another way, growth per month during the school year could be determined by dividing the growth from one April 15 to the next by 10 (not 12).

The California Achievement Test, 1970 (CAT/70) is an example of a test that uses this 10 month year assumption in its norms tables. Even the April 15 "empirical" norm data on this test is actually a projection (using the above assumption) from the normative data that was collected in February, March, and April. The methods of extrapolation and interpolation previously discussed are equally applicable for year-to-year testing except that growth estimates over the school year follow the "10 month year" assumption. For example, consider a grade 2 Title I reading project with pretest scores from near the end of first grade (April 15) on the CAT/70 Level 1, Reading Total, and posttest scores from the same test administered on June 1 of the following year. Assume that the mean April 15 pretest standard score was 209 and the mean posttest score on June 1 of the next year was 272. Assume further that the testing was delayed until June 1 in one school because of local logistical problems and that there was a desire to compare results in that school with

other schools that tested on April 15. A linear interpolation over the 13.5 months between April 15, 1976 and June 1, 1977 would not make any allowance for the presumably slower growth rate over the summer. Following the "10 month year" assumption there would then be 11.5 months from April 15, 1976 to June 1, 1977; dividing the gain of 63 standard score units by 11.5 yields a growth rate of slightly less than 5.5 standard score units per month, or a gain of about 8.2 units in the 1.5 months from April 15 to June 1. Subtracting the 8.2 from 272 yields a predicted standard score of 263.8 on April 15 which corresponds to a percentile rank of 10.6 (NCE = 23.7). If growth were assumed to be linear over 12 months (instead of 10) the gain of 63 standard score units would be divided by 13.5, for a gain of 4.7 standard score units per month or 7.0 units from April 15 to June 1. This yields a predicted standard score of 265 on April 15 (percentile rank = 11; NCE = 24.2). As long as the actual test date is within six weeks of the midpoint of the empirical norming period it should make little difference whether a 10 or 12 month school year is assumed.

If it were desired to determine the percentile rank of the Title I group as of June 1, the extrapolation of the norms could follow essentially the same procedure used in the fall to spring testing, except that monthly gains for the norm group would again be determined by assuming a 10 month year.

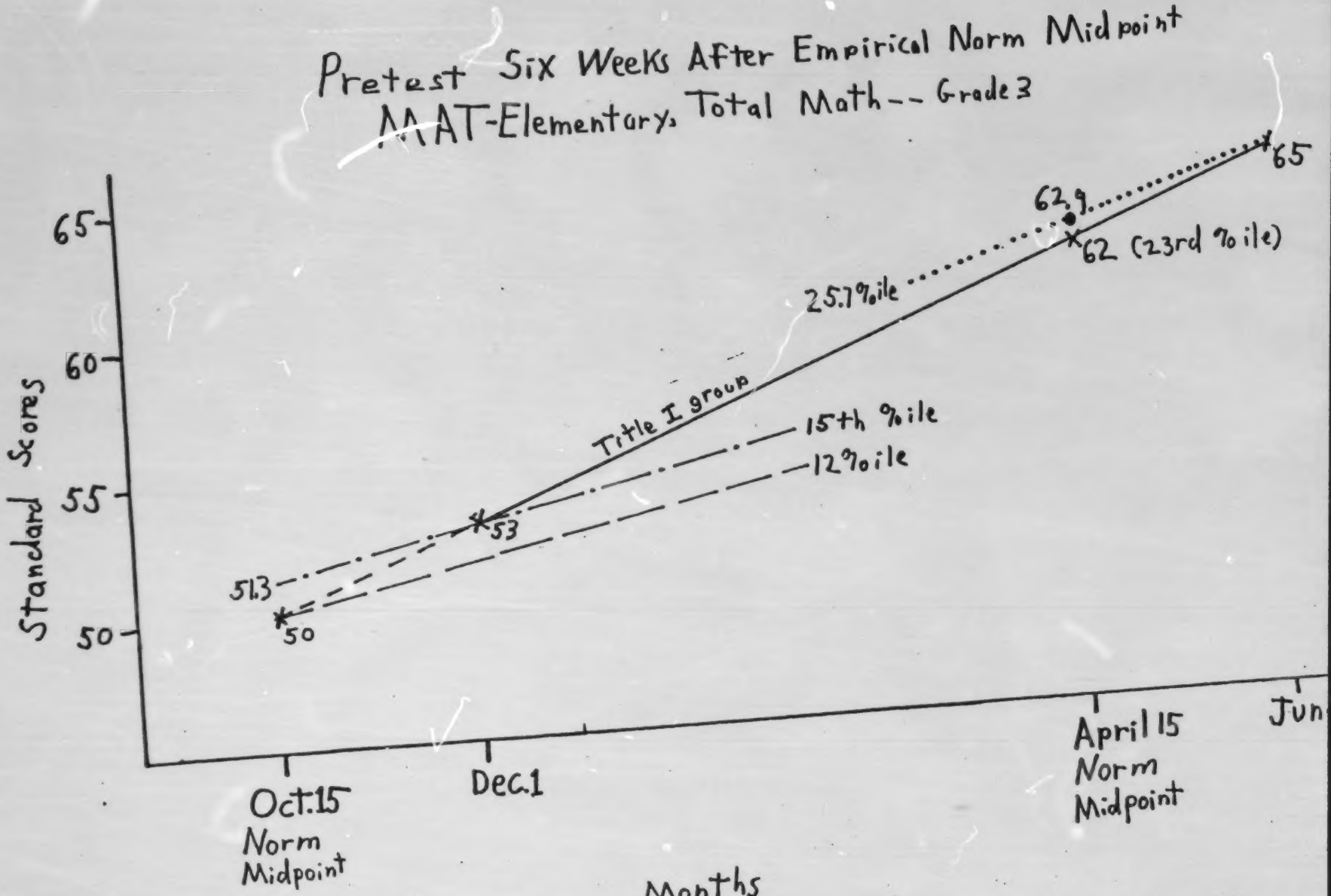
## Conclusions

The procedures outlined in this paper should provide reasonable estimates of program impact when testing dates do not coincide with the empirical norm midpoints. Since the number of months between pre- and posttests in different projects could vary considerably, and assumptions used in extrapolations or interpolating could also differ, comparisons of the amount of gain among various projects must be made with extreme caution.

## Reference

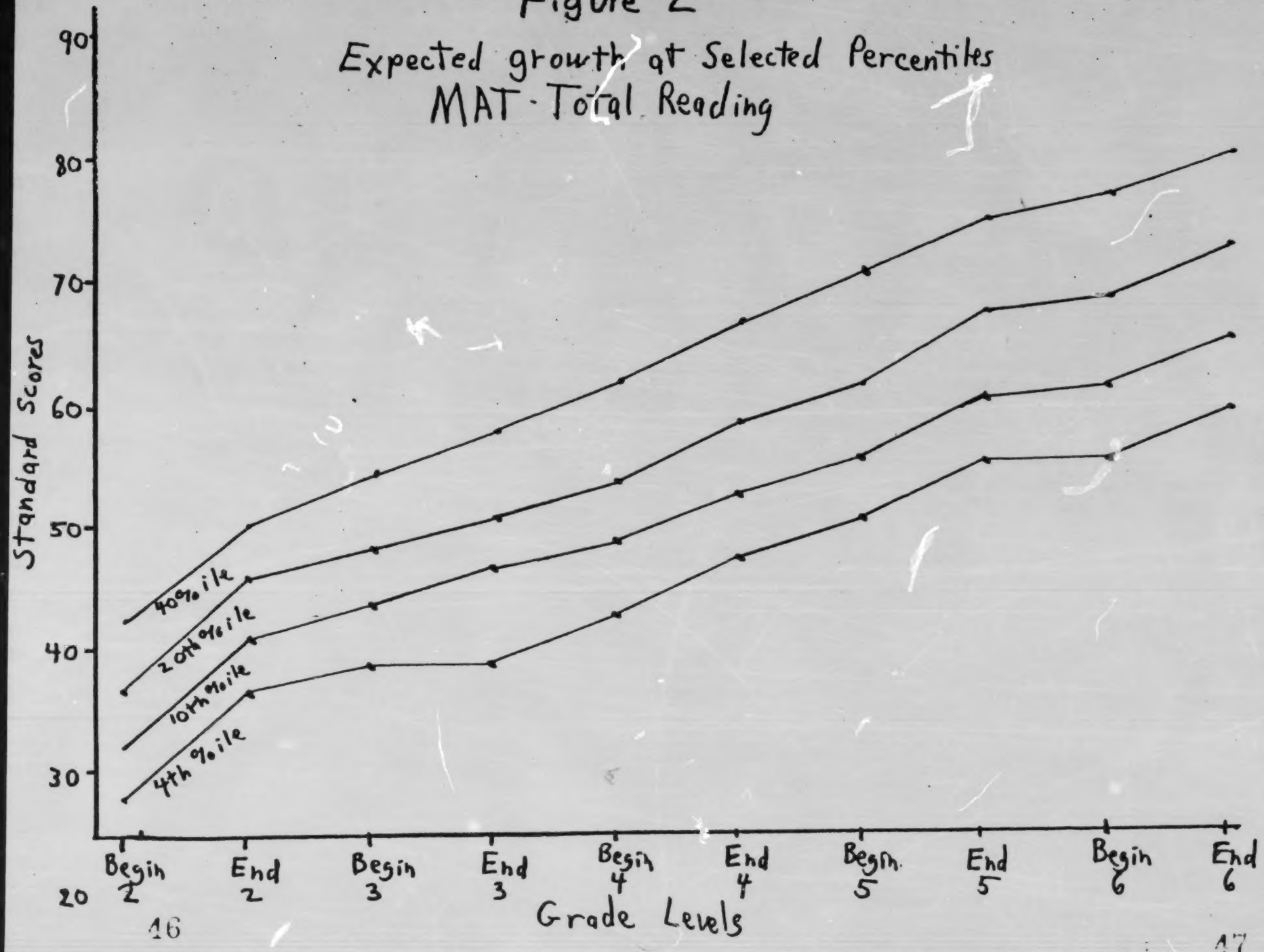
Tallmadge, G. K. and Wood, C. T. User's Guide: ESEA Title I Evaluation and Reporting System. Mountain View, California: RMC Research Corporation, 1976.

# Figure 1

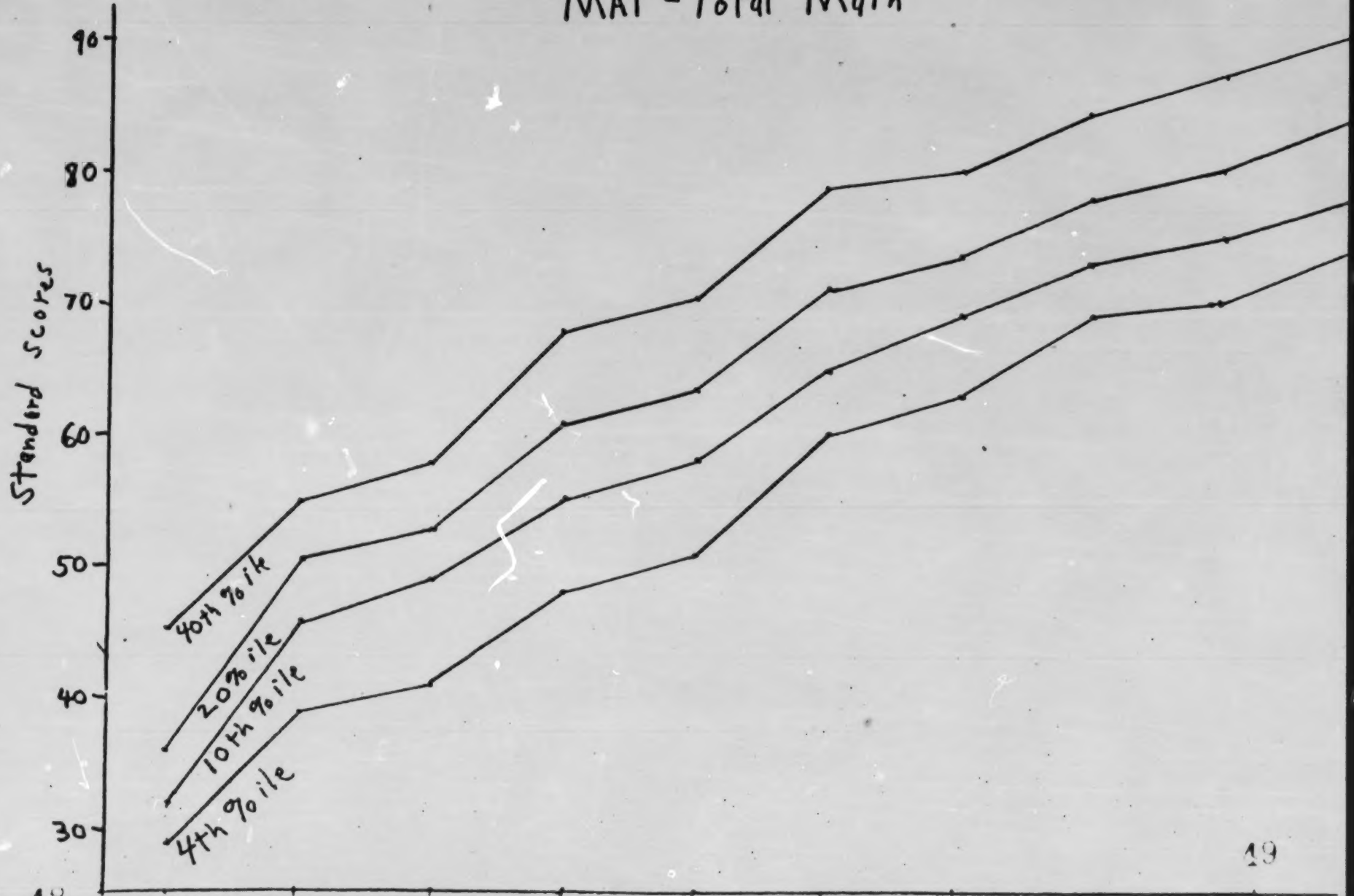

Pretest Six Weeks After Empirical Norm Midpoint
MAT-Elementary, Total Math -- Grade 3
Standard Scores
65
60
55
50
62.9
65
62 (23rd %ile)
25.7%ile
Title I group
15th %ile
12%ile
53
51.3
50
April 15
Norm
Midpoint
Jun
Oct. 15
Norm
Midpoint
Dec. 1
Months

# Figure 2

## Expected growth at Selected Percentiles
## MAT - Total Reading



# Figure 3
## Expected Growth at Selected Percentiles
## MAT - Total Math



nts NATIONAL TESTING SERVICE, INC.
1905 CHAPEL HILL ROAD / DURHAM, NORTH CAROLINA 27707 / TELEPHONE AREA CODE (919) 493-3451

# TECHNICAL COMMITTEE RECOMMENDATIONS REGARDING TESTING TIMES IN ALL MODELS

Robin Baker, Ph.D., Keith McNeil, Ph.D.

It was the unanimous opinion of the Technical Advisory Committee that all testing, both pre and post, should be done as near to the norming dates as possible. In this way the information that is collected can be used not only for the OE evaluation models but can also be used at the individual student level in computing valid norm-referenced scores.

In Models B and C, however, it has been stated that testing does not necessarily have to be done at the norming date. This was motivated by the fact that Models B and C have local control or comparison groups, while Model A depends on the national norming sample as a control group. Since information about the growth of the national norming sample is valid only at the norming dates, Model A necessitates testing at the norming dates. The question has arisen in Models B and C that if one is not testing at the norming dates, should NCEs or expanded standard scores be used to report the individual student achievement levels? It was the opinion of the committee that if one is testing within six weeks of the norming date then interpolation should be used to convert the individual student score to an NCE. Since the actual testing date is relatively close to the norming date, any errors that might arise from interpolation are likely to be extremely small. On the other hand, if



Presented at Project Director's Meeting, January 16-18, 1978, Austin, Texas

the testing date is far removed from the norming date, say a spring testing with a test that only has fall norms, then it was the committee's recommendation that expanded standard scores should be used. It was felt that to convert raw student scores to NCEs using the "nearest" available norms would be inappropriate. While we can make claims about NCEs being equal interval at the norming dates themselves, such assumptions are extremely tenuous when we are testing away from the norming dates.* This is because for any given level of a test, the distributions of scores will change depending upon the grade level. Since NCEs are totally dependent upon the distribution, to assume that the NCEs at one norming point are the same NCEs that one would get at a subsequent norming point is questionable. Expanded standard scores, however, reflect these changes in distribution. Consequently, it was the committee's opinion that the use of standard scores would be more sensible for Model B or C when testing is far removed from the norming date. Standard scores should be used for all calculations up to and including the gain. The gain expressed in standard scores would then be converted to an NCE gain using an appropriate linear transformation based on the standard deviations of the expanded standard scores and the NCEs.

It should be noted that the committee's recommendations regarding whether to use NCEs or expanded standard scores depending on testing date is in no way related to the question of whether NCEs or standard scores should be used when the Title I group is scoring at an extremely low percentile rank.

*An example was computed to determine the difference in treatment effect in three different situations: 1) only fall norms used for both pre and post; 2) only spring norms used for both pre and post; and, 3) fall norms used for pre and spring norms used for post. The differences in treatment effect were dramatic.

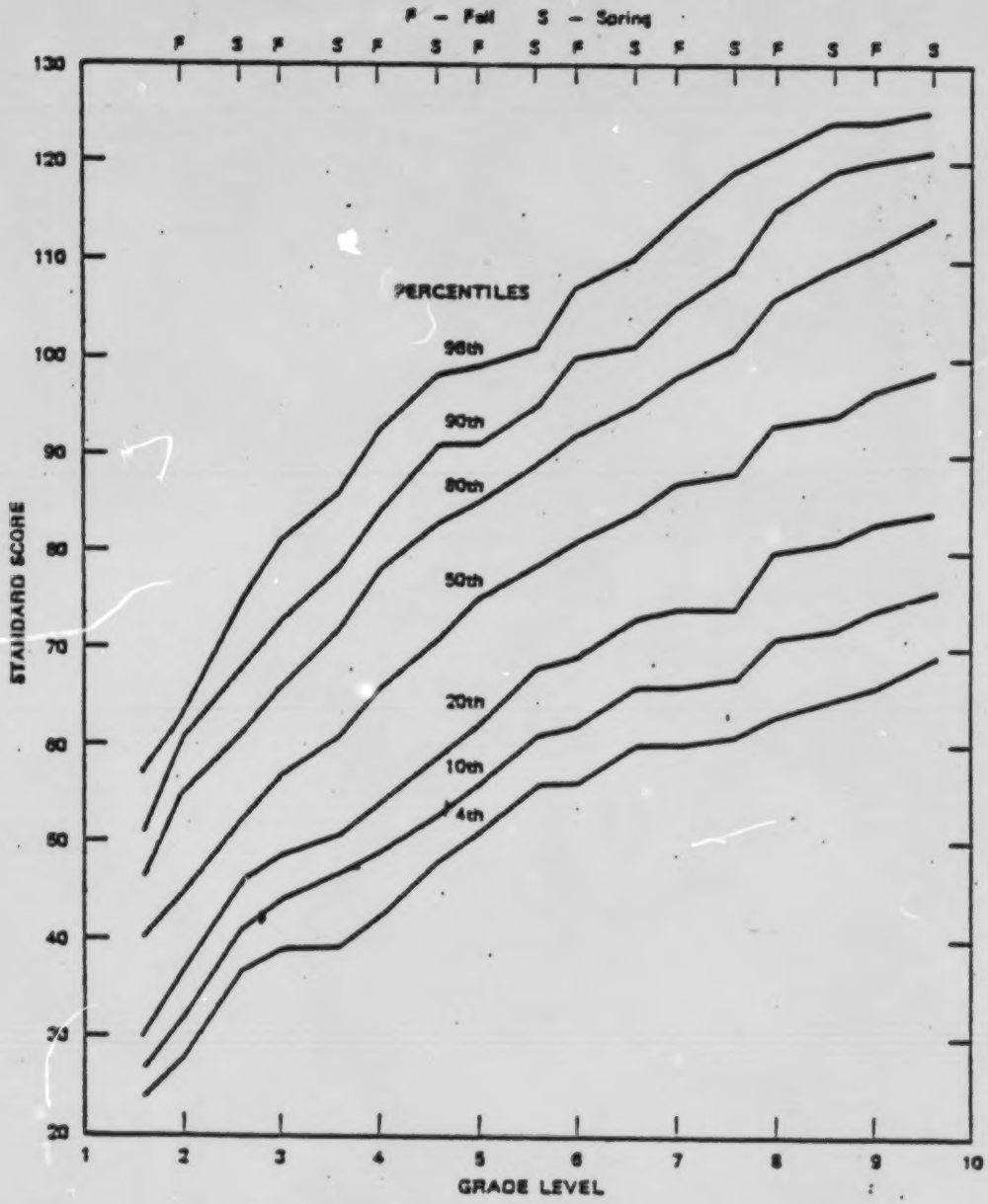



Figure 2.1 STANDARD SCORES BY GRADE LEVEL FOR SELECTED PERCENTILE RANKS: MAT TOTAL READING

# ISSUES RELATED TO INTERPOLATION

Robin Baker
Tom Williams

NTS/TAC

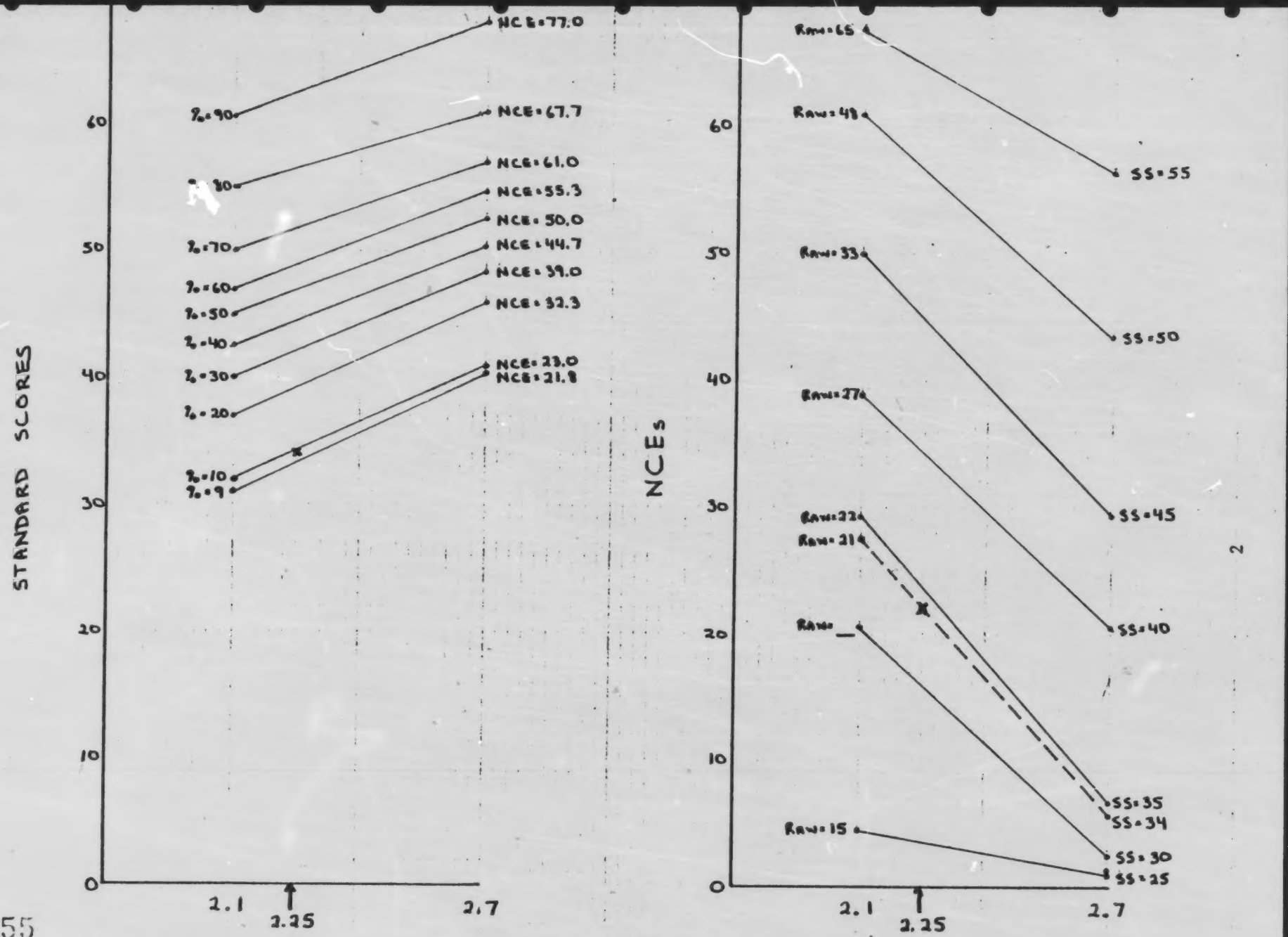
## ISSUE I: PROCEDURES FOR PROJECTION

The standard procedure for projecting normative data can be described as follows: a graph is drawn where time (using a 40 week year) is the x-axis and standard scores are the y-axis. A family of 99 line segments is placed on this graph where the endpoints for each line segment are the standard scores necessary to be placed at a certain percentile rank on the norming dates that bracket the local testing date. An example of such a graph is presented in the left-hand portion of figure 1.1. The data are based on the Metropolitan Achievement Test (MAT) grade 2 norms for the Total Reading subscale . An "x" has been used to represent a standard score of 34 achieved in a November 30/December 1 testing. The assumption is made that since standard scores are equal interval, growth between the two norming dates, if expressed in standard score units, can be assumed to be linear. Thus, each line segment is an equi-percentile line which we assume represents the standard scores necessary to maintain the same percentile rank for all time points between the norming dates. Then, to find the percentile rank for a given standard score from the local testing, a point is placed in the graph at the appropriate location for the standard score and testing date. The percentiles of the equi-percentile lines that bracket the point are identified, (in our example, the 9th and 10th percentiles) and the percentile rank of the point is computed based on its relative distance between these two lines. It occurs to us that in the OE system, we should actually be calling them equi-NCE lines (although, they're exactly the same lines) and calculating the NCE of the point rather than its percentile rank. This is because we assume that NCEs are equal interval (while percentile ranks are not) and thus suitable for any calculations based on relative distance. Then to find the NCE in our example, we interpolate the standard scores at the local testing date for the 10th and 9th percentile lines (34.25 and 33.39, respectively), and then calculate that the NCE must be

$$21.8 + (23.0 - 21.8) * \frac{34.0 - 33.39}{34.25 - 33.39} = 22.65$$

Tallmadge has taken this one step further and developed a procedure that should be much quicker than the one we've just described. Instead of using equi-NCE lines, we use equi-standard score (or equi-raw score) lines on a graph which again has time as the x-axis but has NCEs as the y-axis. An example of such a graph is presented in the right-hand portion of Figure 1.1. Thus, we assume that each line represents the NCEs that a certain standard score produces for all time points between the norming dates. We assume that no growth (maintenance of the same standard score) is reflected by a linear deterioration in NCEs.

Tallmadge's procedure will be faster since only the line corresponding to the observed standard score needs to be computed. With the former procedure, at least two lines that will bracket the observed standard score need to be computed. This will probably involve calculating several lines until the appropriate two lines are found. Then,




55

56

FIGURE 1.1

one additional calculation is necessary to determine the NCE of the standard scored based on its relative distance between the two lines. For Tallmadge's procedure, however, it is only necessary to interpolate the NCE at the local testing date for the appropriate standard score line. For our example, only the dotted line corresponding to the standard score of 34 was necessary; the other lines were included merely for illustration. The Fall NCE for a standard score of 34 is 27.2; the Spring NCE is 5.6. Interpolation to the testing date yields 21.8 as the NCE.

The question then becomes, are the two procedures identical? Although we have not yet developed a formal proof, it seems intuitively plausible that they are, <u>given</u> that the standard scores and NCEs are linearly related to one another. In other words, the standard scores need to be normally distributed within the grade group at both norming dates. To the extent that they are non-normal, the results of these two procedures will differ. This area could definitely use some empirical investigation.

Which one to choose should, of course, be based on which kind of line most accurately reflects what is occurring across time. Put another way, since we assume that growth is linear, is growth as represented by standard scores more likely to be linear than growth as represented by NCEs? Unfortunately, intellectual growth cannot be directly measured or observed, so the question is moot. We can further assume, however, that growth on a scale will be linear only if the scale itself is equal-interval. NCEs are by definition equal-interval at the norming dates, but we can only assume that they would be equal-interval during the intervening period. This may be a perfectly valid assumption: if a scale is equal interval at one point in time, why shouldn't it continue to be equal interval at other points in time? Score distributions do change across time, however. The scores for one level of a test will tend to be negatively skewed if administered below level, roughly symmetrical if on level, and positively skewed if above level. NCEs are distribution-dependent; consequently, to say that NCEs are equal-interval at a point removed from the norming date is to make the questionable assumption that then when we convert a raw score (or standard score) to an NCE we will only be off some fixed additive constant that is uniform across the range of possible scores. Put another way, the NCE difference between a person getting a raw score of 22 and a person getting a raw score of 27 should be unchanged across time. In Figure 1.1, the lines in the right-hand graph should be parallel. To the degree that they are not is an indication of changes taking place across time that render use of NCEs at other than the norming points inappropriate.

On the other hand, standard scores are built by taking scores that are normally distributed at the norming dates and adjusting them so that they interlock with one another. It can be argued, therefore, that they are based much more directly on growth from one norming date to the next and thus take the changing nature of the distributions into account. This would suggest that the standard interpolation/extrapolation procedure would be the procedure of choice.

This needs to be tempered with a practical consideration, however. The differences in results between the two procedures may not be substantial; as previously stated, this needs further investigation. Given that the differences are small, the computational ease of the Tallmadge procedure would definitely make it the one to use.

## ISSUE II: INTERPOLATION VS. EXTRAPOLATION

We are concerned about what we feel is the use of inappropiate norming points in the interpolation procedure. Typically, the norming points chosen bracket the actual testing date. When the test has October and April norming dates, for instance, interpolation for a grade 6 September testing would be based on the grade 5 April and grade 6 October norms. This assumes that the summer growth rate is the same as the school year growth rate. This may be a totally fallacious assumption. Kaskowitz and Norwood (1977) examined the norms of the MAT in the course of assessing its standardization procedure. They presented the graph shown in Figure 2.1. This graph shows selected equi-percentile standard score growth curves for the total reading subscale of the MAT. It is quite clear from this graph that at least for the MAT norms, the assumption of equal growth rates throughout the year is not being met. Third graders, for instance, at the the lower percentiles show a much slower growth rate during the school year than during the preceding summer. Sixth graders in the lower percentiles, on the other hand, show a faster rate of growth during the school year than during the preceding summer. The standard scores necessary to maintain the twentieth percentile rank in the fifth and sixth grades are shown below:

| Grade | Standard Scores |
|---|---|
| 5.1 | 62 |
| 5.7 | 68 |
| 6.1 | 69 |
| 6.7 | 73 |

If we were to interpolate the standard score necessary to be at the twentieth percentile on total reading at the beginning of September (which is six weeks prior to the midpoint of the Fall norming period), we would calculate a score of 68.7. On the other hand, if we were to extrapolate the score using the norming information from the sixth grade school year then we would get a standard score of 68. This standard score is lower than the one that we calculated from interpolation. The implication of this is that if we were using Model A, extrapolation would give us a somewhat lower no treatment expectation than would using interpolation.

We are arguing that it would seem appropriate that the interpolation/extrapolation procedure should be based solely on within school norms. If the sixth grade is tested, then only grade six norms should be used. If the norms reflect a "Summer drop-off" or "Summer slow down" among low achieving students, then using the standard interpolation procedure will produce a no treatment expectation in Model A that is too high.

Obviously, this whole discussion is moot when the test is normed only once a year. In addition, if the publisher's fall/spring norms do tend to be linear (i.e., the growth is uniform), then we do not have a problem. To the extent, however, that differences in summer and school year growth rates are reflected in the norms, the Model A treatment effect will be biased.

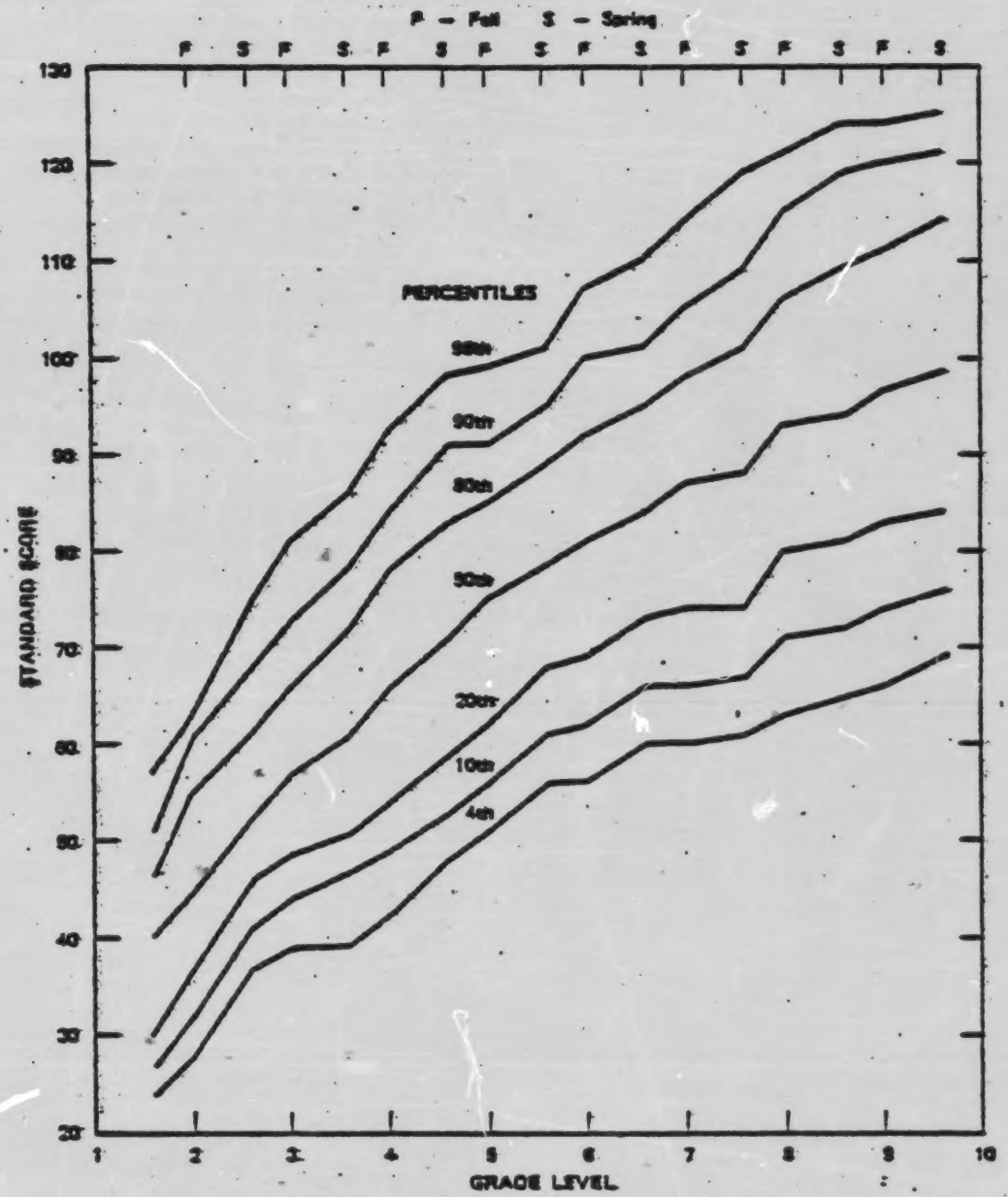



Figure 2.1 STANDARD SCORES BY GRADE LEVEL FOR SELECTED PERCENTILE RANKS: MAT TOTAL READING

References

Kaskowitz, D.H. & Norwood, C.R., "A Study of the Norm-Referenced Procedure for Evaluating Project Effectiveness as Applied in the Evaluation of Project Information Packages." a research memorandum prepared for the U.S. Office of Education, OEC-0-74-9256 (Stanford Research Institute, Menlo Park, California, January, 1977).

Attachment 2

# IDENTIFYING AND CONTROLLING SOURCES OF REGRESSION BIAS IN MODEL A

Stephen L. Murray

Discussion DRAFT Only

Comments Welcome
Do Not Cite

**PREPARED BY:**

**Title I Evaluation**
**Technical Assistance Center**
**Northwest Regional Educational Laboratory**
**710 S.W. Second Avenue**
**Portland, Oregon 97204**

Telephone: (503) 248-6971

## IDENTIFYING AND CONTROLLING SOURCES OF REGRESSION BIAS IN MODEL A

Each of the Title I evaluation models developed by RMC carries certain assumptions which are critical to generating comparable and unbiased Treatment Effect estimates. The equal percentile growth assumption is such an assumption for Model A. If a group of students are not expected to maintain their percentile status with respect to a norm group from pre- to posttest without special treatment, then the no-treatment expectation generated by Model A will be biased. Biased results would mean that estimates based upon Model A could not be aggregated and comparisons between schools using different models would be inappropriate.

Since Model A is the most frequently adopted model, it is important to critically examine the equal percentile growth assumption. The following questions may be addressed: Under what conditions can we expect the equal percentile growth assumption to be valid? How frequently can we expect those conditions to exist? What controls can be exercised to see that those conditions do exist? Can an unbiased estimate of the no-treatment expectation be generated when the equal percentile growth assumption does not seem tenable?

It will be demonstrated in the following argument that the conditions which bear on the validity of the equal percentile growth assumption for a particular group of students include:

(a) The subpopulation from which students are drawn

(b) The expected growth rate of that population in relation to a national population and in the absence of treatment

(c) The process for selecting students from the subpopulation into the treatment group

Work on this argument began by addressing the issue of whether independent administrations of the selection test and the pretest would control for regression effects in Model A when the mean pretest of the selected group is used to define the no-treatment expectation. It was concluded then that implementing Model A may result in a biased Treatment Effect. A key indication of bias is a higher correlation between the selection measure and the pretest than between the selection measure and the posttest: $r_{X_0X_1} > r_{X_0X_2}$.

Where:

$X_0$ denotes selection test scores obtained at occasion 0,

$X_1$ denotes pretest scores obtained at occasion 1, and

$X_2$ denotes posttest scores obtained at occasion 2.

Bias due to regression will result when $r_{X_0X_1} > r_{X_0X_2}$, even though the selection test scores are obtained independently of the pretest test scores. The direction of this bias will result in an overestimate of the Treatment Effect.

This paper extends the conclusion that bias in the Treatment Effect is a function of diminishing correlations across time, a phenomenon sometimes called temporal erosion. The precise function is defined using a formal model adapted from Kenny (1975). The analysis is conducted for the nonequivalent control group design with pre- and posttesting where achievement or some other measured characteristic obtained prior to the pretest is used for selection.

Kenny (1975) has advised that:

--the researcher carefully study the process of selection into treatments. The critical questions are (a) With what and by how much does the treatment correlate with the causes of the dependent variable?, and (b) When does selection take place? The answer to these questions determines the mode of analysis. (page 359)

Accordingly, the selection process is modeled as a causal system which allows us to derive the null hypothesis appropriate for a given selection process. The null hypothesis can be expressed in terms of the expected point biserial correlation between the treatment (coded as a dummy variable where treatment group members are scored as 1 and control group members are coded as 0) and the posttest score ($r_{x_2T}$). The point biserial correlation is a standardized measure of the mean posttest difference between the treatment group and the control group. As such, it is a measure of Treatment Effect. A positive Treatment Effect point biserial correlation indicates a mean difference advantage for the treatment group and a negative correlation indicates a mean difference advantage in favor of the control group.

When a true experiment has been implemented, then our expectation is that the correlation of pretest scores with treatment ($r_{x_1T}$) is equal to 0. If this expectation is met in a particular case, then the null hypothesis expressed in terms of $r_{x_1T}$ would be:

$$H_o: \ r_{x_1T} = r_{x_2T} = 0.$$

The problem arises when $r_{x_1T} \neq 0$, which may be the result of chance or a nonrandom selection process. When this happens, then some adjustment is usually employed to generate the appropriate null hypothesis.

Kenny (1975) has expressed the null hypothesis associated with four alternative adjustment procedures. They are presented in Table 1.

TABLE 1

Null Hypotheses Corresponding to Four Statistical Adjustment Methods

| ADJUSTMENT METHOD | NULL HYPOTHESIS |
|---|---|
| ANCOVA | $r_{x_2T} = r_{x_1T}r_{x_1x_2}$ |
| ANCOVA with reliability correction | $r_{x_2T} = r_{x_1T}r_{x_1x_2}/r_{x_1x_1}$ |
| Raw change score analysis | $r_{x_2T} = r_{x_1T}\sigma_{x_1}/\sigma_{x_2}$ |
| Standardized change score analysis | $r_{x_2T} = r_{x_1T}$ |

Kenny then specified three alternative selection processes and derived the appropriate null hypothesis for each of them. These selection processes and their appropriate null hypotheses are presented in Table 2.

TABLE 2

Null Hypotheses Corresponding to Four Selection Processes

| SELECTION PROCESS | NULL HYPOTHESIS |
|---|---|
| Selection based on the pretest | $r_{x_2T} = r_{x_1T}r_{x_1x_2}$ |
| Selection based on pretest true score | $r_{x_2T} = r_{x_1T}r_{x_1x_2}/r_{x_1x_1}$ |
| Selection based only on group differences | $r_{x_2T} = r_{x_1T}$ |
| Selection between the pretest and posttest | $r_{x_2T} = r_{x_1T}$ |

The selection model which matches the implementation rules for Model A when an achievement test or some other measured characteristic is used for selection might be called "selection based on a measure taken prior to the pretest."

There are four measured variables in the model. They are:

$X_0$ = a measure used for selection,

$X_1$ = a pretest measure of achievement,

$X_2$ = a posttest measure of achievement,

T = the treatment variable.

The causal function of X is based on three unmeasured hypothetical constructs:

G = group differences such as sex, race, classroom, etc.,

Z = individual differences within groups such as ability, initial achievement, etc.

E = errors of measurement or unstable causes of X.

Assuming all variables have been standardized to reduce algebraic complexity, we can express $X_1$ and $X_2$ in terms of their causes, G, Z and E:

$$X_{0i} = a_0 G_i + b_0 Z_{0i} + e_0 E_{0i} \quad (1)$$

$$X_{1i} = a_1 G_i + b_1 Z_{1i} + e_1 E_{1i} \quad (2)$$

$$X_{2i} = a_2 G_i + b_2 Z_{2i} + e_2 E_{2i} \quad (3)$$

where the subscripts 0, 1 and 2 refer to occasions and subscript i to the subject. Note that, since we are deriving the no-treatment expectation, no effects on $X_{2i}$ are assumed to result from T, the treatment variable.

The following assumptions are made for this series of linear equations:

(1) Group membership (G) is assumed to remain constant across time; its autocorrelation equals 1.00.

(2) Relative position within groups (Z) may vary across time; its autocorrelation may be less than 1.00.

(3) Errors of measurement (E) are perfectly unstable, or random; its autocorrelation is 0.

(4) All unmeasured variables (G, Z and E) are assumed to be uncorrelated with each other.

(5) Group membership (G) and relative position within groups (Z) do not interact.

If the treatment is correlated with the selection test, then it must be confounded with the causes of the scores on the selection test. In other words, whether an individual is in the treatment group or the control group is a function of his group membership, his relative status within the group, errors of measurement, or some combination of the three variables. Given this, the causal function for the treatment variable is written as:

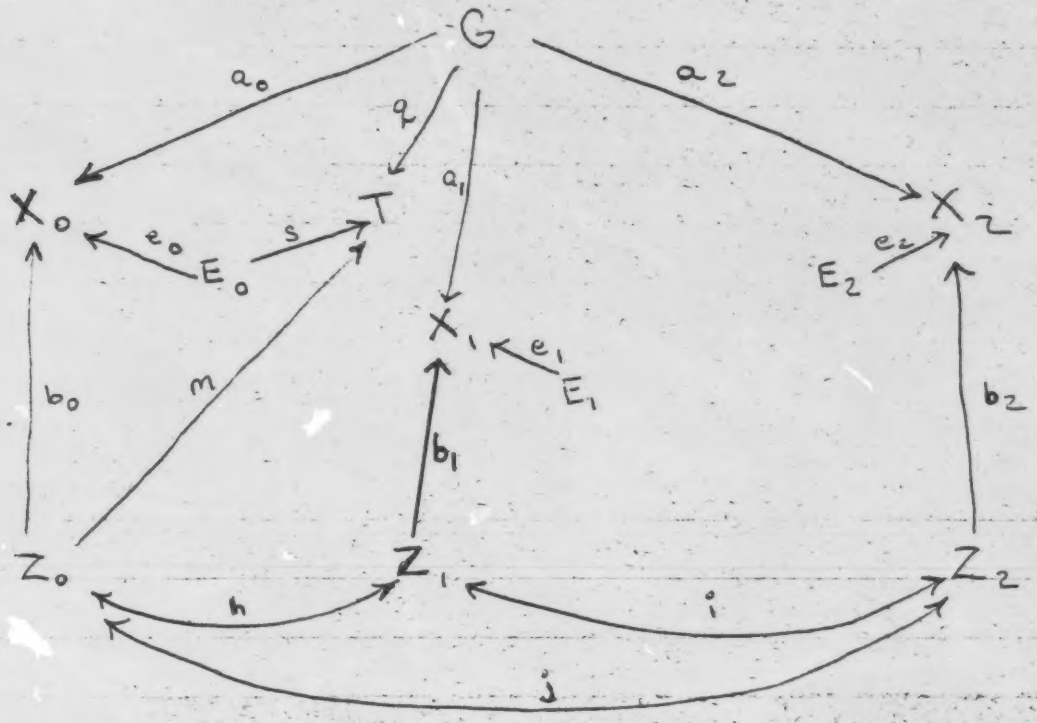
$$T_i = qG_i + mZ_{0i} + sE_{0i} + fU; \qquad (4)$$

where U is a residual term that is uncorrelated with all unmeasured variables. The smaller the residual term, the more reasonable the model that shows treatment and control group membership to be based on the same variables that determine pretest status. A large residual term would indicate that assignment to treatment is based upon variables that are uncorrelated with G, Z and E. Since these variables are unmeasured, a direct assessment of the residual term is not made.

The next step in developing the appropriate null hypothesis is to express the correlations between the measured variables $X_0$, $X_1$, $X_2$ and T in terms of standardized partial regression coefficients. These regression coefficients have been introduced in equations 1, 2 and 3, and they are depicted in Figure 1.

FIGURE 1

Path Diagram Depicting the Application of Independently Administered Tests for Selection ($X_0$) of Students into Treatment (T) and for a Pretest Measure ($X_1$)




Since G, Z and E are latent variables, the regression coefficients linking those variables with the measured variables $X_0$, T, $X_1$ and $X_2$ are also unobservable. However, by making certain assumptions and taking advantage of algebraic identities, we can use these coefficients to show theoretically-expected relationships among the measured variables.

The correlations between the observed variables are expressed in terms of the regression coefficients (path coefficients) in Figure 1.

$$r_{X_0T} = qa_0 + mb_0 + se_0 \quad (5)$$

$$r_{X_1T} = qa_1 + mhb_1 \quad (6)$$

$$r_{X_2T} = qa_2 + mjb_2 \quad (7)$$

$$r_{X_0X_1} = a_0a_1 + b_0b_1h_x \quad (8)$$

$$r_{X_0X_2} = a_0a_2 + b_0b_2j \quad (9)$$

$$r_{X_1X_2} = a_1a_2 + b_1b_2i \quad (10)$$

In each of these equations, the correlation between two variables is taken as the sum of the products of certain regression coefficients.

If selection is based upon $X_0$, we have:

$$T_i = kX_{0i} + fUi, \quad (11)$$

which expresses membership in the treatment group as a function of the selection test score ($X_0$) and a residual (U).

Next, we can substitute equation (1) for $X_{0i}$ in equation (11), which gives:

$$T_i = k(a_0G_i + b_0Z_{0i} + e_0E_{0i}) + fUi.$$

Since T also equals $qG_i + mZ_{0i} + sE_{0i} + fUi$ from equation (4), we have:

$$a_0k = q \;,\; b_0k = m \;,\; e_0k = s$$

or

$$q/a_0 = m/b_0 = s/e_0 = k.$$

Recalling that $a^2_0 + b^2_0 + e^2_0 = 1$ because $X_1$ is standardized, it is possible to show $r_{x_0T}$ is equal to K:

$$r_{x_0T} = qa_0 + mb_0 + se_0 \quad \text{(equation 5)}$$

$$r_{x_0T} = a_0ka_0 + b_0kb_0 + e_0ke_0$$

$$r_{x_0T} = a^2_0k + b_0{}^2k + e_0{}^2k$$

$$r_{x_0T} = k$$

Next we show that $r_{x_1T} = kr_{x_1x_2}$ since $q = a_0k$ and $m = b_0k$:

$$r_{x_1T} = a_0ka_1 + b_0khb_1 \quad \text{(equation 6)}$$

$$r_{x_1T} = k\,(a_0a_1 + b_0b_1h)$$

$$r_{x_1T} = kr_{x_0x_1}$$

Now that we have $r_{x_0T} = k$ and $r_{x_1T} = kr_{x_0x_1}$, we can solve for k in $kr_{x_0x_1}$ and find:

$$r_{x_1T} = r_{x_0T} \cdot r_{x_0x_1}$$

This expresses the correlation between the pretest and the treatment in terms of the correlation between the selection test and treatment and the correlation between the selection test and the pretest. Since $r_{x_0x_1}$ will nearly always be less than 1.00, we can expect $r_{x_1T}$ to be less than $r_{x_0T}$. The equation $r_{x_1T} = r_{x_0T} \cdot r_{x_0x_1}$ is identical to the null hypothesis for ANCOVA if $X_1$ were the posttest rather than the pretest. The solution to this point is identical to the solution offered by Kenny (1975), page 350. Since the logic of Model A requires that we solve for $r_{x_2T}$ in terms of $r_{x_1T}$, additional steps are needed to find these results.

We begin by showing that:

$$r_{x_2T} = a_0k_{a_2} + b_0kjb_2,$$

which follows directly from $r_{x_2T} = qa_2 + mjb_2$ and $q = a_0k$ and $m = b_0k$.

Factoring K out of $a_0k_{a_2} + b_0kjb_2$ gives:

$$r_{x_2T} = k\,(a_0a_2 + b_0jb_2).$$

Knowing that:

$$r_{x_0x_2} = a_0a_2 + b_0jb_2 \quad \text{(equation 9)}$$

leads to:

$$r_{x_2T} = kr_{x_0x_2},$$

solving for K gives:

$$r_{x_2T} = r_{x_0T} \cdot r_{x_0x_2};$$

but, since $r_{x_0T} = r_{x_1T} / r_{x_0x_1}$, because $r_{x_1T} = r_{x_0T} \cdot r_{x_0x_1}$, we have:

$$r_{x_2T} = r_{x_1T}\,(r_{x_0x_2} / r_{x_0x_1}).$$

This final expression gives the expected value of $r_{x_2T}$, where:

(1) A selection test has been used to assign students to treatment;

(2) An independently administered test has been used as a pretest, which is then used to generate a baseline or no-treatment expectation.

The ratio $r_{x_0x_2} / r_{x_0x_1}$ is the appropriate statistical adjustment for $r_{x_1T}$, assuming that regression is homogeneous and linear.

Discussion . . .

It is commonly known among educational researchers that regression effects inevitably accompany imperfect test-retest correlations for groups selected for their extremity and tested again. Low- or high-scoring groups chosen because of their scores

will regress toward the mean of the overall population from which they were selected. Such regression, if not controlled experimentally or statistically, will bias estimates of treatment effect. The confounding of regression effects with treatment effects can make programs look more effective or less effective than they actually are.

The norm-referenced evaluation model (Model A) developed by RMC attempts to control regression effects by requiring that the measure used as the pretest not be used for selecting students into the Title I program. Often, this means that an achievement test other than the one used for the pretest is used for selection. The mean pretest score of the treatment group expressed as a percentile equivalent is defined as the no-treatment expectation. When the pretest and the selection measure have a higher correlation than the selection measure and the posttest, regression effects will still be confounded with the treatment effect. The result will be an overestimate of the treatment effect of the Title I project.

The theoretical model used to derive the adjustment in this working paper specifies that group membership is fixed across time and relative status within groups is left free to vary from perfect stability. The argument that $r_{x_2T} = r_{x_1T}\, r_{x_0x_2}/r_{x_0x_1}$ is not invalidated if this assumption is invalid. We may have specified only that $r_{E_0E_1} = r_{E_0E_2} = r_{E_1E_2} = 0$. 'Temporal erosion' may result from group instability or the instability of individuals within groups. The model was used primarily to maintain clearly the relationship between the correction developed here and the other corrections derived by Kenny.

Perhaps the key to determining whether or not Model A as recommended will result in a regression bias is knowledge of the selection process. Granted, there is an

empirical question whether $r_{x_0x_2}$ is less than $r_{x_0x_1}$ (the real question is probably how much less $r_{x_0x_2}$ is than $r_{x_0x_1}$) but, a post-hoc examination of these correlations may be too late, or even neglected.

When measured achievement is the sole basis of selection, less control of bias is provided than would be provided by the use of the selection test as the pretest with a correction factor equivalent to ANCOVA.

Recent work on quasi experimentation stresses that failure to randomly assign subjects to treatment and control groups does not necessarily leave one unable to estimate treatment effects free from bias. Where the selection process is known to the extent that it can be mathematically modeled, it is still possible, in theory, to provide unbiased estimates of treatment effects (Cain, 1977; Rubin, 1977; Kenny, 1975; Bryk and Weisberg, 1977)--although Cronbach, Rogosa, Ploden and Price (1976) point out that it is quite difficult to effectively control the parameters which do affect bias.

Even accepting the more optimistic views of the potential for control through selection, the selection rules for Model A are not clear enough to warrant this kind of certainty. In fact, recommending any student measure be used to select students into Title I programs (a practice which has merit in providing scarce resources to those most in need) makes the assumption behind the Model A no-treatment expectation less believable. The validity of the Model A no-treatment expectation is based on the blanket acceptance of the assumption that selection into the Title I program is made solely on the basis of stable group constructs.

The selection of students into Title I programs may vary from the model presented here. For instance, students may be selected first because of their membership in a particular group whose mean achievement remains stationary. If there is no further

selection, or random selection only is used to assign students to treatments, then the proper null hypothesis is $r_{x_2T} = r_{x_1T}$. When selection is complete, or random, within a population, the expected value of $r_{x_1T}$ is 0.

However, where the group is chosen first and measured achievement is used to select students from within that group, some correction is warranted as $r_{x_0T}$ is negative by design and both $r_{x_1T}$ and $r_{x_2T}$ are negative due to indirect selection. Local norms or an unselected group would be necessary to estimate the value of $r_{x_1T}$, however.

Practical Conclusions and Recommendations

1. Using any measure as the major determiner of selection, independent of a pretest, violates the key assumption which underlies the Model A no-treatment expectation, namely, that selection into Title I programs is made solely on the basis of stable group constructs (i.e., membership in a definable population).
2. If an achievement measure is used for selection, and is independent of a pretest, it does not necessarily control for regression bias.
3. If Model A is implemented, with an achievement measure being used for selection, an unbiased estimate of the no-treatment expectation will require a statistical adjustment.
4. The appropriate adjustment procedure is based on the relationship between the correlation of selection and posttest measures and the correlation of selection and pretest measures.
5. If $r_{x_0x_2} = r_{x_0x_1}$, or is nearly so, the equal percentile growth assumption will provide a fairly good estimate of the treatment effect. The appropriate statistical adjustment would support the equal percentile growth assumption where it is valid.
6. To apply the correction prior to the posttest, it is advisable to estimate $r_{x_1T}$ from a local norming on an unselected sample.

7. As a departure from the current version of Model A, one could eliminate the separate administration of the selection measure and use the pretest for selection. The no-treatment expectation would be based on the regression of the posttest on the pretest, as in ANCOVA.

8. It is recommended that the argument presented in this working draft be given external review to evaluate its soundness.

9. It is recommended that some TAC(s) develop alternative procedures for modifying Model A, in light of the conclusions presented in this paper.

PSYCHOLOGICAL BULLETIN
A JOURNAL OF THE
AMERICAN PSYCHOLOGICAL ASSOCIATION

DAVID A. KENNY, ASSOCIATE EDITOR
HARVARD UNIVERSITY
33 KIRKLAND STREET
CAMBRIDGE, MASSACHUSETTS 02138
(617) 495-3715

RECEIVED
NWREL
1978
ESEA Title I.T.A.C.

8 March '78

Joe:

Here's the memo for the Denver mtg. Could you see that it's sent to those who'll be there? It's light reading, and they can get them it after they arrive in Denver if it's more convenient to put it in their mailboxes at the hotel.

G²

To: Unknown, Interested Parties
From: Gene Glass
Subject: Regression Effect
Date: 8 March '78

I want to raise the issue of what I regard as a misconception about the workings of the regression effect in pretest-posttest designs. I understand that this misconception has found its way into the Title I Technical Assistance Center materials -- I have this information by word of mouth only since I have not obtained a copy of RMC's booklet no. 3 and I am awaiting a copy of revised materials on the question from Bev Anderson. But the misconception was firmly entrenched at RMC when I consulted with them in 1975 and my antagonistic criticisms did nothing to loose it -- quite the opposite, probably.

The issue is this: RMC may believe, or the TACs may believe or survey or publish the notion that a "double pretest" gets rid of the regression (toward the mean) effect in pretest-posttest designs. The reasoning goes like this: if a selected (for low scores, let's say) group on a pretest will regress toward the mean of

the group from which they were selected on a posttest, why not give them a second pretest immediately, or soon, and let them get all the regressing out of their system; then there should be no regression toward the mean from pretest #2 to the posttest?

A related form of the misconception has turned up in the drafts of the AERA-NCME-APA et al. standards for evaluation:

"Also, the regression problem could have been avoided by administering a separate pretest following the administration of the selection test. In that event, the regression to the mean from the pretest would have been similar in the pretest and posttest."

The second sentence in this quotation is somewhat ambiguous. It could be that the first time the word "pretest" was used in the second sentence that "selection test" is intended. Both of the two possible interpretations of the paragraph are in error, viz., 1) that the amount of regression from "selection test" to "second pretest" equals the amount from "selection test" to "posttest," or 2) that the amount of regression from "selection test" to "second pretest" equals the amount

from "second pretest" to "posttest" so that the former can be used to correct the "second pretest - posttest" difference for the latter, thus rendering an "invalid" design, "valid." Either interpretation is wrong -- or at best, correct only under unusual and impractical circumstances.

Back to the original misconception and the demonstration of why it is an error: consider administering pretest #1, selecting those who score below some cut-off score, then administering pretest #2, then administering a posttest [forget about "treatment effects" for the time being]. Let's assume the conclusion that some would have us believe, i.e., that there should be no regression toward the mean from pretest #2 to posttest, and show that this assumption runs afoul of something we know to be obviously true.

If there is no regression from pretest #2 to the posttest (but of course there is regression from pretest #1 to pretest #2, or we wouldn't have administered the second pretest), then

$$\bar{X}_1 < \bar{X}_2 \quad \text{but} \quad \bar{X}_2 = \bar{X}_3 , \qquad (1)$$

where the subscripts are 1 = pretest #1, 2 = pretest #2 and 3 = posttest. (Assume many cases and no sampling error so that these values are essentially parameters; nothing in the argument to follow depends in any way on this assumption.) The conditions in (1) or those that the advocates of the "second pretest design" would have us believe.

But now, what do we know about the regression effect as an article of unshakable faith? That it is proportional to the correlation of the two variables involved, i.e.,

$$\text{Regression from \#1 to \#2} = r_{12}\left(\frac{\bar{X}_1 - \bar{X}}{s_x}\right) - \bar{X}_1$$

where $\bar{X}$ is the mean of the entire group from which the subgroup was selected.

Now if $\bar{X}_2 = \bar{X}_3$ and the regression of the group from #1 to #2 is the same as the regression from #1 to #3, then

Regression from #1 to #3 $= r_{13}\left(\frac{\bar{X}_1 - \bar{X}}{s_x}\right) - \bar{X}_1$ is

equal to

$$r_{12}\left(\frac{\bar{X}_1 - \bar{X}}{s_x}\right) - \bar{X}_1 . \qquad (2)$$

It's clear that (2) implies that $r_{12} = r_{13}$. But this is the condition that we know immediately to be unrealistic and impractical, namely, that pretest #1 have the same correlation with pretest #2 and the posttest. All the psychometric measures I know of -- and surely those for which this "design" was intended -- correlate more highly the closer together they are in time.

Permit me to put the argument in a somewhat more direct form. Suppose pretest #1 and pretest #2, which are given two weeks apart, correlate .80 -- a reasonable assumption. Suppose that pretest #1 and the posttest, which are given 9 months apart, correlate .50 -- also a reasonable assumption. The regression from $\bar{X}_1$ to $\bar{X}_2$ will be 20% of the distance toward the

total group mean [20% = 100% (1 − .80)]. The regression from $\bar{X}_1$ to $\bar{X}_3$ will be 50% of the distance ~~from~~ toward the total group mean. Hence, $\bar{X}_2$ has to move 30% of the distance toward the ~~mean~~ total group mean by the posttest -- it's impossible for $\bar{X}_2$ to equal $\bar{X}_3$, which equality the "double-pretest design" people wish to believe is so.

I've encountered one final dodge by the "double-pretest design" people in their attempts to salvage what they've invented. They claim that the "double-pretest design" takes out all regression due to "measurement error," or that the design is perfectly valid when all regression is due to measurement error. The reasoning -- antiquated and impractical as it is -- runs like this: measurement errors are uncorrelated, with zero mean and constant variance; if one selects low scores on pretest #1, all of their negative measurement errors will have a chance to average out to zero on pretest #2, thus they'll regress in from pretest #1 to pretest #2 the full extent of any regression of which they are capable -- there'll be no regression left from pretest #2 to the posttest. The miscon-

ception here is somewhat hidden, but rest assured it is present all the same. The misconception stems from the belief that "measurement error" is an intrinsic property of a test instead of a characteristic of a particular use of a test. Suppose I administer Test X twice with 24 hours intervening, and then I administer Test X twice with one year intervening. Will the "one-day reliability" be different from the "one-year reliability"? Obviously, yes. Then the variances of the two sets of measurement errors will differ (since there's no reason to suspect that the observed score variances will differ). Therefore, the measurement errors cannot be an intrinsic property of the test since they are clearly different quantities in these two instances. There are <u>different</u> measurement errors, with different variances, depending on whether one is speaking of one-day reliability or one-year reliability. So although the Double-pretest design takes care of the "errors of meas." between pretest #1 and pretest #2 (as such errors exist over similar intervals of time), it does nothing about the "errors of measurement" that exist for the

interval between pretest #2 and posttest.

Notice that if the "errors of measurement" are intrinsic to a test, then its reliability would be the same whether assessed as stability across one hour, one week, or one year. Hence, $r_{12}$ would equal $r_{13}$ and, as we saw on p. 5, the "double-pretest design" would work. But the errors of measurement are not intrinsic to a test. The history of reliability theory over the past three decades has been the development of the theory that reliability is not a property of a test, but rather a property of a particular use of a test (cf. Cronbach on generalizability theory, for example).

---

In short, the "double pretest design" is bad advice based on some misunderstandings about regression and reliability.

P.S. The afternoon's mail brought news from Bev Anderson in the form of a photocopy of "The Regression Effect" by G. Kasten Tallmadge (January, 1978). It shows that the regression misconception thrives, and it gives some clues as to its origins.

GKT develops a long illustration of the regression effect built on the example of flipping fair coins. He comes to the following conclusion:

"Above-average test scores reflect a predominance of positive measurement error, while below-average scores reflect a predominance of negative measurement error. Furthermore, scores greatly below average reflect more negative error than scores moderately below average.

Coins or groups of students selected because their observed test scores are below average will tend to have below-average true scores. Their observed test scores will also tend to have a predominance of negative measurement error. Although error was randomly distributed across the total group tested, it is necessarily different from zero for low- and high-scoring subgroups.

When a below-average subgroup is retested, the expectation is that error will be randomly distributed across the group members. Negative errors and positive errors will balance each other and the mean error score for the subgroup will have an expected value of zero. With average error at or near zero, the second set of scores will be higher than the first set that reflected a preponderance of negative error. The selected subgroups will thus score closer to the mean of the original group on the second testing than on the first. This phenomenon is called regression toward the mean, statistical regression, the regression artifact or, simply, the regression effect.

If selected subgroups are given a third testing, there will be no further regression toward the mean of the original group. Assuming no treatment, the difference between the second and third sets of scores will have an expected value of zero while both the second and third sets will be closer to the original group's mean than the first set.

Clearly, if the first set of scores are considered pretest measures and the second set posttest meausres there will be an ample gain but this apparent gain will not reflect any real difference in achievement levels. If the first set of scores is used for selection purposes, then it is necessary to retest the selected subgroup. The second set of scores provides an unbiased pretest measure third set can be collected to obtain an unbiased posttest measure.

Selecting low-scoring students for participation in Title I projects is a valid and endorsed procedure. Low-achieveing students will be identified in this manner, although some students not selected will have lower true achievement levels than some who are selected. The observed scores of the selected students will, however, reflect a preponderance of negative measurement error and, for that reason, should never be used as pretest scores in norm-referenced evaluations. <u>The only legitimate procedure is to select the pupils first and then pretest them.</u> If selection is based on test scores (and there is nothing in Models A or B to suggest that it must be), then a second set of test scores must be collected to serve as pretest measures."

This position is wrong for the reasons presented on pages 1-8. And now it is clear how RMC fell into the error. They over-generalized their coin-flipping example. The coin-flipping example just happens to be a special case of "zero reliability" and, furthermore, "zero reliability" independent of elapsed time between the two flips. Hence, $r_{12} = r_{13} = r_{23} = \ldots = 0$.

Concrete examples can facilitate reasoning; they can also confound it.

# A Practical Alternative to Model A

Brent Bridgeman
Region II Title I TAC
Educational Testing Service

As the systematic bias in Model A becomes more widely recognized and better documented (see recent paper by Stephen Murry, memo from Gene Glass, and simulation by Gary Echternacht), it is increasingly important to find unbiased alternatives that would be acceptable to LEAs.

First, Model C could probably be sold much more effectively than it has been up until now. Many LEAs and SEAs rejected this model without giving full consideration to its possible advantages (e.g., the possibility of testing only once a year). The primary objection to Model C appears to be a lack of the necessary computing power at the local level. But this apparent disadvantage often disappears under closer scrutiny. Some states (e.g., Iowa) have noted that even the supposedly simple procedures in Model A are not carried out accurately at the local level, and require centralized computer processing. If computer processing is needed anyway, why not use Model C? It is important to emphasize that LEAs need not have their own computing facilities or programming staffs; they need only send their scores in a prespecified format to some central location. Start-up costs (to get the program running properly) could be shared with numerous other LEAs who were sending their scores to the same central location. Operational computing costs would be minimal; total costs including keypunching and analysis would still be lower and possibly much lower than the costs involved in the additional testing required under

current Model A rules. A number of SEAs should be able to provide the central computing facilities themselves, while those SEAs without their own computers could facilitate the establishment of Model C computing centers at one or two local universities or with private contractors.

Another objection to Model C is that additional students must be tested. But all school systems that get Title I funds must have a procedure for selecting the students with the greatest educational need from among the Title I-eligible students. Whether this procedure is a standardized achievement test or something else, it should still be correlated with posttest achievement scores and hence serve as a legitimate Model C pretest.

With centralized computer processing most LEAs should be able to use Model C, but some alternative model is still needed for the LEAs where the comparison group is too small to create a reasonably stable regression line. One simple (unfortunately overly simple) method of handling the regression problem in Model A would be to select students with the pretest and estimate the regression effect from stability coefficients furnished by the test publisher. Thus, for example, if the pretest and posttest correlated .70, pretest scores would be expected to regress 30% (i.e., 1-.70) of the distance to the mean. For a group with a pretest mean NCE of 30, the posttest expectation would be that the group would regress 30% of the 20 points (50-30=20) that it was initially below the mean. The no-treatment expectation would then be 36. This would be the best estimate for a random sample of the population with pretest scores of 30. But we know that Title I-eligible students with NCEs of 30 are not

a random sample of all students in the country with NCEs of 30. The Title I-eligible students differ from other students in the norm sample on certain relatively stable group characteristics (e.g., socio-economic status) that are known to be correlated with achievement test scores. Since Title I-eligible students have an NCE mean that is lower than 50 they would be expected to regress toward this lower mean and not the mean for the entire population. Thus, if the Title I-eligible mean were 35, the no-treatment expectation for our Title I group with a pretest mean of 30 would be 31.5 [i.e., .3(35-30)+30] and not 36. Even this estimate would not be correct since the pretest-posttest correlation in the Title I-eligible subpopulation would also be different (and possibly substantially different) from the population value of .7.[1]

In some instances the pretest-posttest regression line for the appropriate subpopulation and the appropriate pretest-posttest time interval may be found in previously published data from the numerous studies of compensatory education projects. Since stability coefficients for the major published tests are so similar it should even be possible to use NCE units to estimate the extent of regression on Test X from scores on Test Y. LEAs would not have to actually compute any regression lines, they would simply enter a table with the pretest scores and find the posttest expectation. Of course, these estimates would still represent some gross aggregations based on national studies. The Title I evaluation system itself could be managed to provide much better estimates for particular LEAs. The regional centers for Model C evaluations mentioned above would already have estimates of the appropriate regression lines

[1]Note that these same arguments apply to a Model C evaluation. The regression line from the comparison group should be based on Title I-eligible students, or students as much like Title I-eligible students as possible. A regression line based on higher SES students would have a different slope and intercept.

based on all the districts that used Model C. The no-treatment expectation for the "revised Model A" districts would then come from a pooled estimate based on the two or three Model C districts that were most similar to it demographically, and ideally that used the same tests. These estimates based on Title I schools from demographically matched districts in the same region of the country should be much more accurate than expectations based simply on national norm groups.

# A CASE FOR INSTRUCTIONALLY DATED NORMS

by Henry Scherich

The purpose of this paper is to make a suggestion for test publishers and test users regarding the way empirical norm dates are derived.

In 1976, when RMC Research Corporation published its recommendations regarding the use of standardized tests in Title I evaluation, one of the important recommendations was the use of tests which had empirical norming dates near the time of the evaluation testing (Tallmadge and Wood, USER's Guide: ESEA Title I Evaluation and Reporting System. U.S. Department of Health, Education, and Welfare--Office of Education, October 1976). At the same time, RMC reviewed eight of the major achievement test batteries and found that most (five of eight) were only normed once during the school year. All the publishers provided norms tables for two or three times per school year, but only three had empirical data from two standardizations (Tallmadge and Wood, Characteristics of Eight Commonly Used Nationally Normed Tests, U.S. Department of Health, Education, and Welfare--Office of Education, October, 1976). The other norm tables were developed by interpolation and extrapolation.

Since then, several publishers have developed additional norms for their tests by using data from their scoring service. In addition, as new tests are being developed, the publishers are norming them more than once during the school year. (For examples, see NCME Measurement News; New Tests. Early Winter, 1977, Volume 20-4). While this seems to be a step in the right direction and the effort of the publishers is to be applauded, one additional bit of information about the way norms are described would be helpful.

The standardization plan for a test generally permits the schools to administer it to the standardization population any time between two dates.

For example, in a fall standardization, the schools may be asked to test between October 3 and October 21. If this is accomplished, the mid-point of the norm period is October 12. However, the number of days of instruction that the students of the standardization schools have received at the time of the testing may vary a great deal. The norms which are obtained will probably accurately reflect the average achievement of the norming group on October 12, but it may not provide the user with a relevant empirical norming date.

The reason this may be a problem is that while most states require about the same number of days of instruction during a school year, the beginning dates and the number and lengths of vacations vary greatly. Using the fall standardization example cited above, students in some early-starting school districts may have had 40 days of instruction before the testing, while in some late-starting school districts, the testing date might have been only a few instructional days into the year. In addition, there could have been teacher strikes, weather interruptions (especially in spring testing), and other local factors, which would increase the discrepancy even more. Finally, there are a few school systems that have 12-month programs for which a norming date would be totally inappropriate. The calendar-dated, empirical norm does not allow for these factors.

Therefore, it would be appropriate to suggest to publishers that they ask standardization schools to test after (or near) a certain number of instructional days. Testing according to the instructional days will negate the effect of starting date, strikes, and other variables on the standardization. The testing would yield an assessment of the achievement at a certain point in the instructional year, and it provides norms that are not tied to the calendar year.

Instructionally dated norms should be much more useful for the users than calendar-dated norms. For example, if the publishers would indicate that the test had norming dates approximately 30 and 150 days into the school year, it would allow the user to administer and interpret the test results in keeping with the district's calendar; not the publisher's. It would also be much more meaningful to interpolate/extrapolate on either side of an instructional year norm.

1/4/78

## References

Kaskowitz, D.H. & Norwood, C.R., "A Study of the Norm-Referenced Procedure for Evaluating Project Effectiveness as Applied in the Evaluation of Project Information Packages." This is a research memorandum prepared for the U.S. Office of Education, OEC-0-74-9256 (Standard Research Institute, Menlo Park, California, January, 1977).

Attachment 3

MEMORANDUM

July 11, 1978

TO: Joe

FROM: Steve

SUBJECT: EQUIPERCENTILE COMMITTEE RECOMMENDATIONS

Comments on points (4) and (5) . . .

A. I agree that there is currently no mathematical model for which the "Model" A-1 assumption holds.

B. There already is some theoretical work in developing models for which scores have a multivariate normal distribution. Both Gary Echternacht and I have developed alternative models.

C. To ask for a model which specifies no regression from test administration 2 to test administration 3 is not necessary. We already have such a model; it is the reigning Model A. We also have models that Gary and I have developed, and both of them can be easily interpreted to show the conditions under which the equipercentile assumption would be expected to hold. Our models indicate that it is an empirical question, depending on how the treatment group was selected and the stability of the test scores involved. Model A ignores the selection process and applies the equipercentile assumption as an accepted general truth. In other words, RMC has said the validity of equipercentile assumption is not an empirical question, and Gary and I have said it is.

I would revise point (4) to read:

> There is currently no mathematical model for which the "Model" A-1 assumption holds. An evaluation of alternative theoretical models, which may include those already developed by TAC personnel, should be conducted. These models should be compared to the reigning version of Model A, which is essentially a verbal theory with no underlying statistical rationale. These theoretical models should make clear the conditions under which regression would not be expected to bias treatment effect estimates.

I would also revise point (5) to read:

> Work on mathematical models should be conducted to take into consideration the problem of relating Title I treatment groups to an appropriate population, or populations. It should consider whether various subpopulations (especially Title I) can be expected to have normally distributed NCE scores, and should describe the implications of the various possibilities. It should critically examine the sense in which Title I students are not reasonably seen as members of a "national population" and describe the implications of these findings.

SLM:gbs
cc: Beverly Anderson
Gary Estes
Barbara Williams